doi: 10.1038/nature08819

Supplementary Note 1. Sequencing, assembly and annotation.

Sequencing. The zebra finch DNA for shotgun sequencing, and for bacterial artificial chromosome (BAC) and cosmid libraries was derived from a single male (Black 17) domesticated zebra finch from the laboratory of Dr. Arthur P. Arnold in the Department of Physiological *Science* at UCLA, Los Angeles, CA, USA. The parents of this male hatched in the same clutch in an aviary of group-housed zebra finches, and therefore may have been brother-sister.

High molecular weight genomic DNA was prepared from the muscle of Black 17. The frozen tissue sample was placed on dry ice and sliced with a fine blade as small as possible, and immediately suspended in 100 mM Tris-HCl (pH 8.0), 100 mM EDTA (pH 8.0), and 100 mM NaCl. The suspension was incubated in the presence of 0.5% sodium dodecyl sulfate (SDS), 100 μg/ml proteinase K at 50^{o}C overnight. The DNA was extracted successively with phenol saturated with TE [10 mM Tris-HCl (pH 8.0), 1mM EDTA] and chloroform. The DNA was ethanol precipitated and dissolved in TE. A BAC library was constructed from the same bird at the Clemson University Genomics Institute (http://www.genome.clemson.edu/) and is available for purchase.

Assembly. The initial assembly was generated using PCAP[1] from ~6X coverage in whole-genome shotgun reads, a combination of plasmid, fosmid and BAC-end read pairs. The sequences of 35 finished BAC clones were incorporated into the final assembly. Assembly statistics demonstrate similar findings to the published chicken draft assembly[2] (Supplementary Table 1). For example, the N50 statistic, defined as the largest length *L* such that 50% of all nucleotides are contained in contigs of size *L* was 36 kb and 39 kb (n=8,037) for chicken and zebra finch, respectively. For supercontigs we produced N50 values of 7 and 10 Mb (n=29) for chicken and zebra finch, respectively. Contigs are contiguous sequences not interrupted by gaps whereas supercontigs are ordered and oriented with estimated gap sizes. The zebra finch BAC physical map contains 108,725 clones for a ~10X depth of coverage and is contained in 2,724 contigs. Map construction was done as previously described[3].

We assessed the base-level accuracy, structural integrity and coverage of the zebra finch genome assembly by comparison to finished bacterial artificial chromosomes (BACs), ESTs[4, 5, 6], and the chicken genome. Over 98.1% of the assembly is composed of bases of PHRED quality 20 or greater[7]. Comparison of the WGS sequence to 51 finished zebra finch BACs with a total length of just over 7 Mb revealed a high-quality discrepancy rate $5x10^{-4}$ and $1x10^{-4}$ insertion/deletion errors, not unexpected given the estimate of the heterozygosity rate in this individual of 1 in 630 as 50% of the polymorphic alleles in the WGS sequence will differ from the single-haplotype BACs.

The structural accuracy of the genome assembly is high based on the comparisons to finished BACs from the sequenced individual as well as based on comparisons to ESTs. Based on the BAC alignments, some small supercontigs (most <5kb) have not been positioned within large supercontigs (<1 event per 100kb). While these are not strictly errors, they do affect the utility of the assembly. There are also small undetected overlaps (most <1kb) between consecutive contigs (~3 events per 100kb) and additional small contigs incorrectly inserted within larger supercontigs (~0.7 events per 100kb). No misoriented or misordered contigs within supercontigs were found. Thus the nucleotide-level accuracy of the assembly is high and the assembly correctly places the vast majority of the zebra finch genome in long contiguous stretches. Comparisons to the ESTIMA EST set [4] revealed no assembly errors for 28,863 ESTs that aligned to the assembly.

We estimated coverage of the finch genome using the alignments to the finished BACs, EST sets, and finch markers[8]. Based on alignments to finished BACs, we estimate we have 94% coverage of the finch genome. Of those BACs, 87% are aligned over more than 90% of their length. Similarly, over 93% of a combined set of 86,275 ESTs from the Duke[6], ESTIMA (http://titan.biotec.uiuc.edu/cgi-bin/ESTWebsite/estima_start?seqSet=songbird3)[4] and Rockefeller[5] databases aligned to the current zebra finch genome. Alignments against the subset of 31,658 ESTs in the ESTIMA set revealed at least partial alignment for 96% of the ESTs. Alignments to the 28,005 assembled cDNAs generated from six tissue-specific cDNA libraries as a part of this project revealed coverage for a majority of the cDNAs, 93%. Comparison of the chicken RefSeq set to the zebra finch genome revealed similar coverage of 93%. Finally, of the 860 zebra finch markers generated for a genome-wide linkage map[8], only 14 (1.6%) did not have an alignment in the zebra finch genome.

Using SSAHA2[9] (http://www.sanger.ac.uk/Software/analysis/SSAHA2/), zebra finch sequencing reads generated by the consortium were aligned to the contigs associated with the taeGut3.2.4 assembly of the zebra finch genome.

The aligned reads are analysed by the pile-up package included with SSAHA2, which takes as its input the full SSAHA2 alignments of all of the reads and determines SNPs by creating a multiple alignment of all of the reads and the reference assembly and uses this structure to determine SNPs at each position. The pile-up pipeline assigns a confidence score to each SNP based on the read mapping score and the PHRED score for the variant base.

To ensure quality and because we are sampling only two zebra finch chromosomes, we require the identification of both a high-confidence reference base and a high confidence variant base at each SNP location. We define high-confidence as having a SNP confidence score greater than 10 and a PHRED score for the variant base greater than 23.

The above procedure results in a set of 1,748,362 SNPs across the zebra finch genome that equates to an observed SNP every 705 bases. The SNP sequences are available in dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/).

Of the 1.2 Gb genome, 1.0 Gb was ordered and oriented along 33 zebra finch chromosomes and 3 linkage groups (Accessory Table 1). The zebra finch chromosomes were named based on their homologous chromosomes in *Gallus gallus*. For those chromosomes where multiple zebra finch chromosomes correspond to a single chicken chromosome, a letter was appended to the chromosome name[10]. Accessory Table 1 lists the cross-referencing of the *Gallus gallus* homologous names with the current naming convention for the zebra finch. All unanchored supercontigs have been concatenated into chromosome "chrUn", separated by gaps of 25 bp. On all other chromosomes, unknown gap sizes between supercontigs have been set to 100 bp.

To create chromosomal sequences, data from the zebra finch genetic linkage map[8] and the physical map were integrated with the whole genome shotgun (WGS) assembly data. Zebra finch SNP marker sequences were assigned by BLAST alignment to contigs in the WGS assembly. Based on these marker assignments, the supercontigs were assigned to a chromosome based on a majority rule (>50% of markers assigned to the same chromosome). The supercontigs were initially positioned along chromosomes based on their median marker position, and initially oriented based on relative marker order along the supercontig. The physical map was also linked to the sequence assembly by using BAC end sequence links and *in silico* digests of the assembly to create "ultracontigs", ordered/oriented lists of "supercontigs". Following these initial placements, the WGS assembly read pairing data were used to aid in orientation and to confirm order. For the Z chromosome, marker order was also determined by fluorescent *in situ* hybridization (FISH; Art Arnold, personal communication) and integrated again with the linkage map, physical map and assembly. All discrepancies between the various maps were manually reviewed and a combined super/ultracontig order was established based on reconciling the data from the genetic linkage map[8], physical map and assembly. Available EST data[4-6] and alignments with the chicken genome were also examined and used as aid in orientation particularly when other zebra finch-specific data were inconclusive. The location of the centromere was known only for the Z chromosome at the time of assembly order and orientation. Thus no other centromeres were placed in the current chromosomal assemblies.

Gene annotation. The Ensembl[11] gene set for zebra finch was augmented by gene predictions arising from the Gpipe pipeline[12]. The two predicted gene sets were compared and additional gene predictions from Gpipe were validated following orthology assignment using the OPTIC pipeline[13] and then by manual inspection. This process provided 472 additional protein-coding genes to the Ensembl gene set. Total RNA extracted from zebra finch spleen, muscle, skin, liver, testes and embryo was used to construct normalized cDNA libraries[14] and were sequenced on the 454 Titanium instrument (Roche). All samples originated from the University of Sheffield zebra finch colony. RNA was pooled from fourteen 5-day-old embryos for the embryonic analysis, from ten nestlings in the case of skin and from five adult males for the other tissues; the same five males were used for muscle and testis. All reads are available from the NCBI short read trace archive. Using these 454 Titanium reads along with zebra finch brain cDNA sequences determined in previous studies[4,5,6], it was possible to extend 3,798 Ensembl genes, with 2,428 extensions located at the 3' end of a gene, 627 extensions at the 5' end and 743 at both ends of a gene.

The Ensembl gene set predictions were annotated using information from the chicken and human genomes available from public resources. All gene symbols and gene descriptions were based on the human gene nomenclature as previously described[15], unless a manual or chicken gene annotation was available. Gene and comparative gene information was downloaded from Ensembl (version 54) using Biomart (http://www.ensembl.org). Gene information from the chicken genome was also downloaded from NCBI (http://www.ncbi.nlm.nih.gov). Human gene and comparative information was downloaded from HGNC and HCOP (http://www.genenames.org). For chicken genes predicted by NCBI but not by Ensembl, we used BLAST to define homologies and these were merged with other gene data above. The first stage of the annotation process defined orthologs between zebra finch, chicken and human gene predictions. Orthologs were defined using information on sequence homology, phylogenetic trees and conservation of synteny. Homologous pairs were downloaded from the Ensembl Compara database (http://www.ensembl.org) together with the orthology type. It was assumed that all 1:1 orthologs were correct and were used to defined conserved syntenic regions. Further orthologs were then defined from the one:many and many:many relationships, if the homologs mapped to a conserved syntenic region. This allowed us to increase the number of defined orthologs for all species by 10-25% (Accessory Table 2). Defined gene information for gene orthologs, including gene symbols and gene descriptions were transferred to the zebra finch gene predictions in order of confidence: manual, chicken and human annotations. Each zebra finch gene record had one of six states: (a) Pending (3,589): gene symbols in progress, where we had no homologs defined as described above (384 have similarities with other genes, which need to be investigated further), (b) Approved (14): official gene symbols that are available from manual annotations, (c) Reserved (0): official gene symbols that are not publicly available, (d) Symbol Withdrawn (0): previous symbols for genes which now have different approved symbols, (e) Entry_Withdrawn (1; ENSTGUG00000009041 artifact looked like *TLR1B*): for symbols which refer to a gene that has since been shown not to exist, and (f) Automatic (14,516): gene symbols assigned automatically. In total, 14,527 zebra finch gene predictions were given a gene symbol. All automated steps were performed using a series of Perl scripts. All gene annotations are available from the Ensembl genome browser.

Brain gene expression. Forebrains of 5 male juvenile (post hatch day 50, PHD50) and six male adult (PHD 850) zebra finches were dissected. Two to three forebrains of each age were pooled resulting in 4 samples (2 x PHD50, 2 x PHD850). Total RNA was extracted from each pooled sample using the

Guanidinium thiocyanate-phenol-chloroform extraction method (Trizol, Invitrogen) according to the manufacturer's protocol.

Library preparation (polyA+ RNA purification, first strand synthesis, removal of dNTPs, second strand synthesis, DNA fragmentation, UNG treatment) and cluster amplification were performed as described[16]. Sequencing was carried out on the Genome Analyser II (Illumina) by performing 51bp paired-end run according to the manufacturer's instructions.

Reads were aligned to the genome using the TopHat software version 1.0.8[17], where reads mapping to genomic regions are reported as well as reads identifying exon-exon junctions. The software was run with the following parameters: maximum intron length 20 kb, minimum intron length 20 bp, mate inner distance of 40 bp (standard deviation 20 bp) and a minimum isoform fraction of 5%. The read alignment track is available here: http://www.molgen.mpg.de/~richard/finch_brain_illumina_mapping.zip

The expression level of a gene (Ensembl version 55) was obtained by counting the number of reads that mapped within its exons. A gene was considered expressed when its read counts were significantly above the value expected for background noise. To model the background noise, we estimated a negative binomial distribution over non-coding regions (Ensembl and NCBI annotation), hypothesizing that read counts scale linearly with the length of a region. The distribution parameters are estimated on regions with a length ranging from 1 to 5 kb. In each replicate, the set of genes declared as expressed is defined to achieve 1% of expected false discovery rate[18]. The differential expression of a gene between the two age categories was assessed with the edgeR tool (http://www.bioconductor.org/packages/release/bioc/html/edgeR.html). A negative binomial distribution was used to model the variability between conditions. The *p*-values computed with this strategy were further adjusted[18]. Genes with an expected false discovery rate below 5% were called as differentially expressed.

Analysis of regulatory regions. The genome was scanned in overlapping windows of length 500 bp (shifts of length 250 bp) for each of 99 distinct motifs (position weight matrixes; PWMs) from the JASPAR database (http://jaspar.cgb.ki.se/). Each window was scored for each motif using an HMM-based score for motif clustering[19]. The 1000 top scoring windows in the genome were considered as motif "target windows". Motif density was calculated as the average over all motifs (Accessory Figure 1), where an individual motif's density is the number of target windows of that motif, as a fraction of all windows. Motif density for a specific category of regions was calculated by considering only windows in that category. Region categories were defined for each window, based on its distance from any gene whose territory is as follows: (i) include the 5 kb upstream region, (ii) include the entire gene itself, (iii) include the region upstream of the gene until half the distance to the next gene, (iv) include the region downstream of the gene until half the distance to the next gene. Note that gene territories of neighboring genes may overlap, if their 5 kb upstream regions overlap. Region categories were defined as follows: "NearUp" and "NearDown": less than 10 kb from gene, upstream and downstream respectively; "MedUp" and "MedDown": between 10 and 50 kb from gene, upstream and downstream respectively; "FarUp" and "FarDown": more than 50 kb from gene, upstream and downstream respectively; "CDS": regions annotated as coding sequence; "Intronic": between two CDS regions.

Our motif target predictions are most enriched in the 10 kb upstream regions of genes (Accessory Figure 1a,b, "NearUp"). The 10 kb downstream regions ("NearDown") are also significantly enriched in motif targets (binding sites), while "far upstream", "FarDown", and "Intronic" are significantly depleted in these targets. Windows that are "MedUp", "MedDown" or "CDS" exhibit the least significant difference from the genome-wide average motif density (Accessory Figure 1b).

We counted the number of target windows of each motif, in each category of a gene's territory (as defined above), and estimated z-scores for these counts based on a permutation test. A z-score reflects the extent to which the count of sites is greater than or less than what is expected for that category of regions. Positive values indicate over-representation. We estimate z-scores as described next and similar to an earlier report[20]. First, each motif's name is assigned as the label for each of its target windows. Each motif's target windows are counted for each region category. Then the labels are permuted (via 10^6 pair-wise swaps) among all target windows (regardless of original motif labels), restricting the permutations to be between windows of similar GC content (within 10%). The permuted labels are then tallied by region category, for each motif separately. The process is repeated 1000 times and a mean and standard deviation for each (motif, region category) pair is computed. Using these values, the original score is converted to a z-score. Thus, the region category preference of a motif is conditional on the overall motif preferences reported above. In other words, if a motif is seen to have a preference for say "near upstream", this is over and above the preference that motifs in general have for "near upstream".

Accessory Figure 2 and Accessory Table 3 indicate that different motifs have strong preferences for various region categories. For example, the FOXL1 motif is strongly enriched in "near upstream" regions (z-score > 8), while the SRF motif is strongly depleted in this region category (z-score < -4). Comparison with a similar study of the human genome[20] reveals 13 cases of agreement between the two studies and only two cases of disagreement. For instance, the enrichment of preference of the FOXL1 motif for "near upstream" regions is consistent between zebra finch and humans[20] (Accessory Table 4).

Supplementary Note 2. Unique features of the zebra finch genome

Comparative mapping. The presence of intrachromosomal rearrangements in the majority of autosomes is consistent with rearrangements demonstrated previously from the linkage map [8] and contrasts with the stable avian karyotype and the high degree of synteny confirmed here. There is strong evidence that ectopic recombination is the major driving force of intra- and interchromosomal rearrangements [21]. Given this common mechanism underlying both types of rearrangements, it is interesting that bird genomes show many intra- but few interchromosomal rearrangements, resulting in a highly conserved karyotype structure[22]. This is likely to be the product of an evolutionary process that minimizes the DNA content (mostly through the number of repetitive sequences) and maximizes the recombination rate of microchromosomes [23]. Through this process the properties (GC content, DNA and repeat content, gene density and recombination rate) of microchromosomes and macrochromosomes have diverged to create distinct chromosome types. We proposed a Fission–Fusion Model of karyotypic evolution based on chromosome rearrangement and minimization of repeat content [24]. The consequence of this model is that mammals with a high repeat content (~50%) will tend to undergo many inter and intra chromosomal rearrangements, in effect scrambling their genomes. In contrast, the low repeat content of birds (~10%) will tend to create fewer opportunities for inter chromosomal rearrangements, as we have found.

In order to identify tentative chromosomal rearrangements between chicken and zebra finch, whole chromosomal sequences of orthologous chromosomes were aligned using the program GenAlyzer[25] with default settings. This analysis identified rearrangements in the majority of chromosomes and can be interrogated in greater detail within the accessory file “Physical mapping table 2009-09-16.xls” and Accessory Figure 3. To validate the bioinformatics results by physical mapping using FISH, we selected 141 chicken and 141 zebra finch BACs with orthologous sequence content covering chicken chromosomes 1-15, 17-28 and Z and their zebra finch orthologues, with special reference to 25 tentative chromosomal rearrangements on chromosomes 1-8 and Z. Chicken sequences were retrieved from Ensembl (http://www.ensembl.org/Gallus_gallus/Info/Index) and aligned against the zebra finch genome (http://genomeold.wustl.edu/tools/blast/) BAC clones containing orthologous sequences were selected from the Wageningen chicken BAC library[26] and the Clemson University Genomics Institute zebra finch BAC library (http://www.genome.clemson.edu/), respectively. The results of the BLASTN alignments and the BAC hybridizations are indicated in the accessory file Physical mapping table 2009-09-16.xls. Isolation of BAC DNA was performed using a commercial Midi preparation kit (Qiagen) and labeling with biotin-16-dUTP or digoxigenin-11-dUTP (Roche Applied Science) was performed by nick translation. FISH proceeded on metaphase chromosomes generated from embryonic fibroblasts and/or peripheral blood, using streptavidin-Cy3 (Amersham) or FITC-anti-digoxigenin (Amersham) to detect the labeled probe and 4',6-diamidino-2-phenylindole (DAPI) to counterstain the chromosomes[27,28]. The signal position was determined as the fractional length from the p terminus, FLpter[29]. Signal positions were measured using the software ImageJ[30].

The BAC sequence alignments and hybridization signals from FISH mapping analyses were congruent (Accessory Figure 3). These findings provide independent, cytogenetic support for the accuracy of the zebra finch sequence assembly.

Z chromosome analysis. For analysis of male to female ratios of gene expression, we used microarray expression data for zebra finch from Tomaszycki et al.[31] and for chicken from Itoh et al.[32] (GEO accession numbers: GSE6843, GSE6844, GSE6856). Data for gene expression were quantile normalized and filtered in the statistical environment R 2.8.0 using the gtools package from R projects (http://www.rproject.org). In each case, expression levels for probes mapping to the same Ensembl gene IDs were averaged.

Chickens have a region of especially high ratios on Zq, and a region of especially low ratios on Zp near 28 Mb. The MHM (male hypermethylated) non-coding RNA is expressed from Zp and is postulated to be involved in local dosage compensation[33]. In contrast, no MHM sequences are found in the zebra finch genome and genes adjacent to the comparable MHM position (Figure 1) show no special cluster of dosage compensation (low M:F ratios). MHM appears not to play a general role in birds.

Gene family expansion and losses. A summary of gene family expansions is found in Supplementary Figure 3. Orthology assignment followed a procedure described previously[12] and recapitulated here. The procedure has four steps: (1) orthologs are predicted between pairs of genomes using PhyOP[34] using a distance metric derived from BLASTP[35] alignments, (2) pairwise orthologs are combined into clusters, (3) sequences within a cluster are multiply aligned using MUSCLE[36], (4) phylogenetic tree topologies are estimated using TreeBeST[37] (http://treesoft.sourceforge.net/treebest.shtml) and clusters are split into orthologous groups using the pufferfish Tetraodon as the outgroup.

For each pair of genomes, each translated transcript was aligned against every other translated transcript using BLASTP[35]. Alignments with E-values exceeding 10^{-5} or covering less than 75% of the smaller sequence were removed. The remaining alignments were weighted according to a normalized bit score $s_{ij} = 1 - ((max[s'_{ij}, s'_{ji}])/min[s'_{ij}, s'_{ji}]$, where s'_{ij} is the bit score for a BLASTP alignment between sequence i and sequence j.

Orthologs between transcripts were then assigned using PhyOP[34], a tree-based orthology assignment procedure. Orthology relationships between transcripts were then translated into orthology relationships between genes. Pairs of orthologous genes were then grouped into clusters in a graph clustering procedure. We constructed a graph with genes as vertices, and vertices are connected if the adjacent genes have been predicted to be orthologous or in-paralogous. Clusters were given by connected components in this graph.

Genes in each cluster were aligned using MUSCLE[36]. Genes with multiple transcripts were collated into a string of non-redundant exons from all transcripts concatenated in sequence. Genes were translated and aligned in amino acid space and afterwards back-translated into nucleotide sequences. Phylogenetic trees were built with TreeBeST providing the established phylogeny for the amniota[38] and using the "-best" option. Incomplete and inconsistent genes confuse the tree building procedure and were eliminated using a heuristic procedure: if two genes from the same species do not overlap in their multiple alignment, then the shorter gene was removed.

Each tree was then split into orthologous groups using Tetraodon sequences as outgroups such that each orthologous group contains only in-paralogs[39] with respect to Tetraodon. Results for orthology assignment are available through the OPTIC[13] server. Simple 1:1 orthologs are of special interest as they describe the shared and conserved gene repertoire among a set of species. The number of simple ortholog sets, which contain exactly one gene per species, decreases with increasing divergence and as more genomes are added (Accessory Table 5).

Additional filtering of zebra finch family expansions was performed. 3,864 genes of 17,475 (22%) of zebra finch genes are located on unplaced chromosomes (chrUn, chr7_random, etc). Unplaced sequence can occur as a consequence of sequencing artifacts that prevent the assembly algorithm from fully collapsing all reads in a region. As a result, the number of gene duplication events might be over-estimated. To arrive at a conservative estimate of duplication events, we computed a second orthology analysis using a filtered zebra finch gene set. The filtered set excluded all duplicated zebra finch genes on unplaced sequence that were more that 97% sequence identical to their closest paralog, except for instances where cDNA data suggested expression of both sequences (see Mapping and assembly gene models below). This procedure conservatively removed duplications that could result from assembly artifacts, but at the same time also removed all instances of very recent duplications. The resulting estimates thus provide a lower bound of gene family expansions in zebra finch. The filter removed 2,519 sequences from the zebra finch gene set. The number of $(1:1)^n$ orthologs across all (n=8) species increased from 4,344 to 4,895 owing to spurious duplications that had been removed (Accessory Table 6).

Synapsin 1 loss. Vertebrate synapsins are encoded by three different genes (*Syn1-3*) and are critical for synapse function, having been linked to schizophrenia and seizures[40,41,42]. The phenotype of synapsin deletions is not easily demonstrated in experimental animals before doing a triple knockout[43]. Showing a synaptic effect of such knockouts requires sophisticated electrophysiological and morphological techniques, since the effects are quite subtle[43,44]. We found that *Syn1* and its flanking genes, which are located on the mammalian X chromosome, are missing in the zebra finch and chicken genomes (Supplementary Table 2).

An initial synteny analysis was performed by searching the chicken genome for chromosomal regions representing fragments of the human X chromosome, using NCBI and Ensembl genome browsers. The analysis was then refined by conducting extensive blast and blat alignments of avian genomes using as queries the mammalian orthologs of *Syn1* and flanking genes on the mammalian X chromosome. To search for possible transcripts representing *Syn1* and flanking genes that may not have been mapped to avian genomes, we conducted extensive tBLASTn searches at NCBI of translated avian EST databases using the predicted human proteins as queries.

At least three distinct synapsin genes are present in fish (teleosts appear to have four synapsin genes due to a duplication of *Syn2*) and frogs, as can be seen by examination of their respective genomes, indicating that the duplication(s) that generated synapsin genes occurred early, before the emergence of birds and arguing against a lack of duplication of synapsin genes in the bird lineage. Furthermore, *Syn1* and its flanking syntenic group are present in lizards, as they can be identified in the partial genomic sequence currently available for the green anole (Supplementary Table 2).

We have also recently obtained evidence of *Syn1* presence in crocodilians, as we have been able to amplify a segment of *Syn1* from genomic DNA (Supplementary Figure 4). Thus, *Syn1* seems to have been present in the archosaurian ancestor to birds and crocodilians. The most parsimonious explanation for these combined observations is that *Syn1* and its syntenic group were lost in dinosaurs or in a more immediate ancestor to modern birds. An intriguing possibility is that this loss occurred in conjunction with the reduction in genome sizes thought to have evolved in the saurischian dinosaur lineage between 230 and 250 million years ago (before this lineage gave rise to the first birds) as suggested by Organ et al.[44].

Olfactory receptor evolution. For many years it was assumed, owing to the relatively small olfactory bulb of some avian species, that birds possess a poor sense of smell[45]. Nevertheless, recent behavioral studies indicate that birds are able to exploit olfactory cues for navigation, foraging and kin recognition[46]. The chicken genome sequence, however, was shown to possess over 500 olfactory receptor- (OR-) like sequences, of which about 78 are functional thereby suggesting that some birds may discriminate diverse odorants[47, 48]. Thereafter, a partial survey revealed that nine other birds from seven orders possess more extensive intact OR gene repertoires[49]. Here we report the discovery of 573 OR-like sequences in the zebra finch genome, of which 215 exhibit an uninterrupted open reading frame, which again suggests that zebra finch extensively exploits its sense of smell. Phylogenetic analysis of the avian OR repertoire and comparison to other vertebrates (Supplementary Figure 5) reveal that the majority (97%) of the zebra finch ORs form a monophyletic clade, arising out of the vertebrate OR family 14 (previously described as OR5U1- or OR5BF1-like ORs). Species-specific and thus recent, gene duplication events are also characteristics of other birds[49]. Interestingly, an expansion of OR family 14 is also seen in platypus, and to a lesser degree in opossum, but not in a sequenced reptile, the green anole, *Anolis carolinensis*[14]

PAK3 expansions. *In situ* hybridization with a *PAK3*-specific riboprobe (from clone CK307206) has revealed that *PAK3* is widely and abundantly expressed throughout the brain of adult male zebra finches, including the nuclei of the song control system (data not shown). We note, however, that there is thus far no evidence of differential regulation of *PAK3* in song control nuclei. We conclude that *PAK3* expression is likely to be of relevance to brain function in general. Of several *PAK3*-like genes identified so far, we have identified several brain ESTs (from the ESTIMA collection) representing two truncated *PAK3*-like paralogues. *In situ* hybridization reveals that one of these genes (from clone CK311654, corresponding to gene model ENSTGUG00000006253) is expressed in a spatial pattern highly reminiscent of that for *PAK3*, with widespread distribution including the song control nuclei and no apparent regional regulation. The second gene (from clone DV950892; this EST cannot be unambiguously assigned to a single Ensembl gene model) shows a more restricted distribution, being primarily expressed in the ventricular zone and in satellite cells of the cerebellar Purkinje cell layer, suggesting a possible relation with proliferative zones of the adult songbird brain. A mapping of 454 Titanium reads from six tissues (Supplementary Notes 1) provided confirmatory evidence of *PAK*3 and these *PAK*3-like genes above in the brain but also revealed the majority are expressed in the testis, suggesting this organ may be a major site of action for this expanded group in songbirds. The gene structure and tissue distribution of *PAK3*-like genes is under investigation.

PHF7 expansion. We aligned amino acid sequences of Ensembl 54 gene models for 19 chicken, 19 zebra finch and five outgroup species using Multiple Alignment using Fast Fourier Transform (MAFFT). Protein Neighbor Joining (distance) and Bootstrap analysis were done through Phylip-3.69 to construct phylogenetic trees. Analyses were run with 1000 replicates, all of which were used to generate a consensus tree. Trees show strong support for a monophyletic group of 17 zebra finch genes and 18 chicken genes. Two zebra finch and one chicken genes cluster closer to the mammalian and fish outgroups (Supplementary Figure 6).

Syntenic mapping of zebra finch to chicken reveals no syntenic regions between any of the copies (see Physical mapping table 2009-09-16.xls). To further ensure that there are no small syntenic regions between chicken and zebra finch copies, we looked at neighboring genes up to 500 kb away. No neighboring genes were found in common between any copies across species.

To further check for the possibility of gene conversion we ran the program GENECONV on the amino acid sequences of all the zebra finch, chicken, and outgroup Ensembl models, the chicken and zebra finch DNA sequences together, and the zebra finch DNA sequences alone. No iteration of the program found global or pairwise fragments for any of the groups, supporting the hypothesis that these are independent expansions rather than the product of gene conversion. However, in order to rule out gene conversion definitively, more work must be done.

Of the 19 zebra finch *PHF7* genes, 17 could be aligned to unique Illumina brain RNA reads (see ‘Brain Gene Expression’ section of Supplementary Notes 1). ESTs from ESTIMA[4] were also mapped to 3/19 of the chicken copies. This is suggestive evidence that some of these gene copies may be functionally expressed in the zebra finch.

Zebra finch protease losses and gains. The degradome is defined as the set of proteases present in an organism[50]. The coining of this term reflects the enormous biological and pathological relevance of proteolysis that pervades virtually every aspect of life, including development, apoptosis, host defense, nutrition, reproduction and central nervous system biology[51]. We expect the study of the zebra finch degradome compared to its chicken and mammalian counterparts will contribute to our understanding of the evolution of this bird and its genomic features. In this analysis we focus our attention on those protease gene loss or expansion events linked to neurological function. We also present a preliminary summary of all proteases examined in the finch that show a loss or gain in multi-species comparisons (Supplementary Table 3).

We used a previously assembled set of genes including curated human[52] and chicken (http://www.ensembl.org) protease sequences. Each protein sequence in this starting set was compared to the genomic sequence of zebra finch using the TBLASTN program of the BLAST suite[50,51,52,53]. Putative orthologues of the starting genes were located with BlastSniffer and curated with GeneTuner (http://degradome.uniovi.es/downloads). To minimize the number of missed protease genes, a composite file with all of the TBLASTN hits sorted by chromosomal location was also generated and inspected. Prediction of chicken orthologues was performed by reciprocal best hit analysis. In those cases where orthology could not be established by this method, phylogenetic studies were conducted. Protein sequences were aligned with ClustalX and manually edited with Genedoc. Then, alignments were bootstrapped 100 times with Seqboot and most parsimonious trees were generated with Protpars, both from PHYLIP (Phylogeny Inference Package, version 3.6).

The zebra finch degradome contains about 460 proteases, which is similar to chicken. 380 of the zebra finch proteases have a reciprocal best hit ortholog in the chicken degradome. By contrast, 80 predicted zebra finch proteases have no clear ortholog in chicken. Most of these proteases belong to large complex families, which may cause some false negatives in this analysis. However, some of these specific zebra finch proteases seem to have arisen from specific duplications.

The zebra finch genome features a zebra finch-specific caspase-3 duplication, although this is not experimentally validated. This event may have important consequences in neural development regardless of the role of this protease in apoptosis. In fact, recent studies have shown that zebra finch caspase-3 plays a dynamic role in song-response habituation, and therefore is involved in learning and memory[54]. It is important to notice that this result requires experimental validation, since tandem duplication can be mimicked by artifacts in the genomic assembly. There are only two zebra finch caspase-3 ESTs in the public databases, and both of them correspond to one of the copies. Interestingly, an independent caspase-3 duplication has been shown in opossum[55].

Notably, the gene encoding the serine protease neurotrypsin (*PRSS12*) is duplicated in zebra finch and in chicken, but not in mammals. Neurotrypsin has been linked to neural development in multiple organisms. A 4 bp deletion in human neurotrypsin mRNA is believed to cause mental retardation[56]. Likewise, a Drosophila melanogaster strain lacking the orthologue of neurotrypsin has been shown to suffer a long-term memory formation defect[57].

Finally, we have found a zebra finch-specific tandem duplication of the gene encoding the aspartyl protease β-secretase 1 (*BACE*). Strikingly, both caspase-3 and *BACE* have been shown to play a role in Aβ peptide accumulation in Alzheimer disease[58]. This suggests that β-secretase activity may be different in zebra finch and other birds.

Major histocompatibility complex (MHC) gene evolution. Manual curation of the genome assembly revealed duplication of MHC Class I and Class IIB genes relative to the chicken. For Class I, we find one locus with a full open reading frame, and at least three putative pseudogenes (sequences with a premature stop codon). For Class IIB, we find four functional loci and at least five apparent pseudogenes. Loci were classified as distinct based on sequence divergence, but many of these loci were represented multiple times in the genome assembly suggesting even more extensive duplication. Presumably as a result of this complex duplication history many MHC genes are placed on ‘Chromosome Un’ in the assembly.

We compared MHC Class I (exons 1-7; not all sequences spanned this entire region) and MHC Class II (exons 2 and 3) genes to chicken MHC loci using MrBayes[59] and a general time reversible model with a proportion of invariant sites and gamma distributed rate variation. We ran MrBayes for one million post burnin generations and checked for convergence. For both Class I and ClassIIB zebra finch loci clustered as a strongly supported monophyletic group relative to chicken sequences (Accessory Figure 4). Analyses using the OPTIC pipeline also support duplication of MHC genes in zebra finch relative to the ancestral amniote.

To further characterize the MHC region we sequenced seven BAC clones containing MHC genes and different BACs FISH mapped to at least two different chromosomes[60]. In addition, the assembly itself places CD1 and Blec1 genes, genes that are MHC linked in chicken, on chromosomes 12 and Z, respectively, in zebra finch. MHC linked genes in the chicken are found on at least four chromosomes in zebra finch.

Searches of EST databases revealed that MHC Class I sequences are expressed in the zebra finch brain as they are in mammals[61].

Repeat element and segmental duplication summary. By independent analyses we find that the zebra finch genome is similar to the chicken in its overall repeat composition (Supplementary Tables 4 and 5). In addition to a RepeatMasker analysis (http://www.repeatmasker.org) the zebra finch genome was also masked by a general *P-Cloud* analysis[62].

P-Cloud examines the repeat structure of a genome based on exact word counts. This method does not perform alignment or similarity searches; thus, the computation time and requirements are substantially reduced[62]. *P-Cloud* can identify repetitive regions of genomes including interspersed repeats, segmental duplications, gene families and pseudogenes. Here, we performed a *P-Cloud* analysis under standard parameter settings. The settings were as follows: 16mers; control parameters: 2, 8, 16, 160, 1600; 80% P-cloud oligonucleotides in 10 oligonucleotide windows.

The false positive rate for this analysis was estimated through simulation of a synthetic genome of equal size. For this approach the sequence was sampled from a Markov chain based on dinucleotide frequencies of the zebra finch genome in 1 Mb windows. Estimates of the repeat-derived fraction (X) were calculated using the formula: $Obs = X(1-FN)+(1-X)*FP$, alternately arranged to solve for X as $X = (Obs-FP)/(1-FP-FN)$; where the observed proportion (Obs) is equal to the non-repeat derived portion (1-X) times that false positive rate (FP), plus the repeat derived portion (X) times the sensitivity rate, one minus the false negative rate (FN).

For the first estimate we conservatively assumed that the analysis did not contain a false negative output. For the second estimate, we obtained a RepeatMasker annotation of the zebra finch genome draft assembly from UCSC (http://genome.ucsc.edu). FN was estimated as the proportion of nucleotides identified as repetitive sequence by RepeatMasker but not by *P-Cloud*.

The estimated proportion of the zebra finch genome that is repeat-derived (ignoring false negatives) is 40.6%, which is quite low compared to mammalian genomes (e.g., 72% for the human genome), but considerably higher than the approximately 9.5% estimate from RepeatMasker. It is possible that the main satellite sequences are unidentified or not in the assembly. Under these conditions, 78.7% of the RepeatMasker annotations were also masked by *P-Clouds*, providing an estimated 21.3% false negative (FN) rate for *P-Clouds* with the given parameter settings. Including this FN rate estimate, along with the estimated false discovery rate in the *P-Cloud*-annotated regions of about 2.7%, a revised estimate of the repeat-derived proportion in the finch genome is 51.86%, and the estimated false discovery rate in the *P-Cloud*-annotated regions is about 2.7%.

We also performed an ERV1 (endogenous retrovirus) specific *P-Cloud* analysis to identify any additional ERV1 derived genome sequences below the lower detection limits of RepeatMasker. To analyze the ERV1 family specifically, we parsed all ERV1 elements present in the current RepeatMasker annotation available from the UCSC server. ERV1-specific *P-Clouds* analyses were run under modified conditions. Here, we used 14 mers (control parameters 2, 2, 4, 40, and 400). Under these conditions we determined the sensitivity as the proportion of nucleotides that were identified as repeats by RepeatMasker and were also annotated by the ERV1-specific *P-Clouds*. The FP rates and estimated true ERV1 nucleotide content were calculated as previously described. Using this approach we recovered approximately 2.5 Mb of additional ERV derived genomic DNA sequences (Supplementary Table 5). ERV derived sequences appear to be an even more significant component of the zebra finch genome than that found in the chicken.

Segmental duplication content of the zebra finch genome was assessed using two different methods: one dependent on the assembly and one based on an assessment of excess depth-of-coverage of WGS sequence data in the genome assembly. A BLAST-based whole genome assembly comparison (WGAC) method[63] was used to identify a total of 198,180 pairwise alignments representing putative duplications, >1 kb in length and with >90% identity. High-copy repeat sequences were initially removed using RepeatMasker (species *Taeniopygia guttata*; initial seed alignments >250 bp and >88%; optimal global alignments including repeats >1 kb and >90% sequence identity). The 198,180 alignments corresponded to 384 Mb of nonredundant sequence and 172,670 alignments (87%) mapped to chrUn or random. Among intrachromosomal alignments where both alignments mapped to a chromosome, 89% (13,583/15,285) were defined as tandem duplicates (both pairs mapping less than 1 Mb).

Since larger, high-identity duplications (>94%) are frequently collapsed within working-draft sequence assemblies[64] or may represent artifactual duplications within an assembly[63], we compared these assembly-based results to a whole-genome shotgun sequence detection (WSSD) database of zebra finch segmental duplications. WSSD identifies regions >10 kb in length with a significant excess of high-quality WGS reads[65] within overlapping 5 kb windows. We established thresholds based on the alignment of WGS reads against 40 unique zebra finch BACs. Our analysis was based on a comparison of 11,683,735 zebra finch WGS reads against 400 kb segments of the taeGut1.0 assembly. 13,523,039 reads were remapped to the assembly based on the following criteria: >94% sequence identity; >200 bp non-RepeatMasked bp and at least 200 bp of PhredQ >30 bp. We excluded regions with repeats of <10% divergence from their consensus and all zebra finch-specific repeat sequences. A total of 16,076 regions corresponding to 44.2 Mb were predicted by this approach. Once again the majority (13,872 or 86.3%) could not be assigned.

In summary, we find that only 8.2 Mb of nonredundant taeGut1 sequence assembly are shared by both methods (WSSD and WGAC), providing a lower-bound estimate of 0.8% for large, highly identical duplications (Accessory Figure 5). The majority (>65% or 846/1,327) of these large duplications (> 20 kb) appear tandemly organized (<1 Mb) within the assembly. If we assume that all 44 Mb of WSSD-predicted segmental duplications correspond to collapsed, high-identity duplications, we estimate an upper bound of 6.5% segmental duplication in the zebra finch genome (44.2-8.2) * 2 + 8.2 = 80.2 / 1,223 Mb). Similar to other published draft assemblies, additional experimental characterization and clone-based sequencing will be necessary to resolve these regions of the genome. Additional details may be found at http://eichlerlab.gs.washington.edu/database.html.

MicroRNA expression and annotation. Two independent projects used RNA sequencing procedures to directly determine zebra finch microRNAs (miRs). Each project using independent next generation sequencing technology to analyze non coding RNAs in auditory forebrain, whole brain and liver. As part of the zebra finch genome sequencing project, the reads derived from these two sources were co-submitted to miRBase (http://www.mirbase.org/), who assigned unified sequence identifiers consistent across the two datasets. The resulting information is given in detail in the accessory file, “Tgu-miRs_2010-02-04.xls”. The information includes the sequences of both stem-loop precursors and mature miRs, as well as genomic map coordinates for the precursor sequences.

Supplementary Note 3. The genomic landscape of vocal communication and learning

Mapping and assembly of gene models. EST and 454 reads were mapped onto the zebra finch genome using GMAP[66] with default options. The resultant matches were filtered by 1) minimum alignment coverage of read (at least 90%, unless unique), 2) at least 50 aligned bases, and 3) uniqueness – no other match of this read had an alignment coverage exceeding 95%. Reads overlapping by at least three bases were assembled into transcript models. Nested reads not sharing sequence were not merged, but assembled into separate transcript models.

Two mapping artifacts were removed: (1) reads extending by more than 10 bases into an intron with canonical splice motifs inferred by another read. Reads which extended less than 10 bases were kept but truncated (see below) and (2) reads suggesting an intron of more than 300 bases without additional support by other reads, while the adjacent exons have strong support from at least two other reads were removed. Such reads often joined adjacent genes.

Small gaps (<10 bp) within transcript models were filled using the genome assembly and intron boundaries were refined to correspond to the three major splice motifs GT/AT, GC/AT, and AT/AC by truncating reads within 10 bases of a putative exon boundary and requiring a read depth differential at the intron/exon boundary. Note these transcript models are not biological transcripts, but rather expressed exons linked by shared reads and ignoring the possibility of alternative transcripts.

GO[67] assignments from Ensembl were assigned to gene territories (genomic segments centered around a single Ensembl gene) and extended by 30 kb on either side. Overlapping territories were resolved at the midpoint. In the rare case of nested genes the territory comprised all nested genes and GO categories of all genes were taken into consideration. The significance overlap of non-coding transcripts with these territories was tested as above using simulations, but restricting the admissible space for the placement of segments to the intergenic portions of gene territories with GO assignments. Multiple testing correction was performed using a false discovery rate (FDR)[18] of 0.05. Both GO and GOSlim categories were tested.

Integration with cDNA microarray data. ESTs including those from the SoNG Initiative microarray[4] and the Dong et al.[68] song responsive data set were mapped onto the zebra finch genome assembly using GMAP[66] with default options (Accessory Table 7). Out of 17,877 ESTs on the SoNG Initiative microarray[4], 16,970 (94%) aligned to the genome assembly and 16,310 (91%) remained after filtering for EST coverage (>90%, at least 50 matching bases) and uniqueness. Overlapping ESTs were assembled into 15,009 transcribed loci (TL).

Transcribed loci were annotated using exons of Ensembl 55 gene models. TL mapping exclusively to Ensembl protein coding exons (including known UTR sequence) were classified as coding (6,600), while those mapping into intronic or intergenic sequence were labeled novel (8,409). TL mapping both to coding and intronic/intergenic sequence were labeled as ambiguous (Accessory Table 8).

We could detect coding potential in only a small minority of novel TL (229 out of 8,409), indicating that the vast majority are indeed not coding. We used the method by Kong et al.[69] to predict coding potential. Briefly, the method uses various features (presence and length of open reading frames, sequence similarity to protein sequences in UNIREF50[70]) to predict coding potential and a Support Vector Machine classifier.

Using simulations similar to those described in Ponjavic et al.[71], we find that *novel* transcripts are depleted in introns as expected (Accessory Table 9). An exception is the set FastDown, which shows no such depletion. Given the size and number of TL in the set FastDown, our method would detect a depletion of 15% at $P < 0.05$.

Gene territory enrichment. Using the previously defined gene territory boundaries (above), we computed the territory length of each gene, and averages for each co-expressed gene set from Dong et al.[68]. The gene length was defined based on gene start and stop site annotations, and the intergenic length was defined as the territory length minus the gene length. The territory lengths of a gene set were compared to those of the complement of that set, using a 2-tailed Wilcoxon Rank Sum test.

The gene set "FastDown" was most enriched for long territories (*p*–value 1.7×10^{-28}), followed by the set "SlowUp" (*p*–value 9.3×10^{-10}; Accessory Figure 6a). The gene set "Group 8" was the set most enriched for short genes (*p*–value 2.1×10^{-8}). Accessory Figure 6b shows that the average gene length correlates positively with the territory length, which implies that an enrichment for long or short territories is due to gene lengths as well as intergenic lengths, not due to one or the other exclusively.

Singing cascade gene analysis (Figure 5)

Singing behavior and RNA sample preparation. A total of 54 adult male zebra finches were individually isolated overnight in sound attenuation chambers and in the morning their brains were dissected either before singing or immediately after singing for 0.5 hour, 1 hour, and every hour up to 7 hours after singing began (n = 6 animals per time point). For the singing groups, we only used birds that sang continuously at 25 bouts or more per 0.5 hours. Following dissection, brain hemispheres were quickly frozen and cryosectioned in the sagittal plane at 10 um onto 'PEN membrane glass slides' for laser capture microdissection using the Arcturus XT Microdissection System (Molecular Devices). Brain sections were sequentially dehydrated in ethanols (70%, 95%, 100%) and then in xylene. The striatal song nucleus Area X was collected from at least 7 brain sections per bird and subsequently processed for total RNA isolation using the Pico pure kit (Molecular Devices). The pallial song nuclei HVC, LMAN, and RA, were prepared for a separate companion publication. After evaluating RNA integrity (1.5-2.0 ng of intact RNA from Area X) on a Bioanalyzer RNA 6000 Pico Chip (Agilent), RNA samples were reverse transcribed and cDNA linearly amplified using the WT Ovation Pico kit (Nugen). Just before amplification, polyadenylated Adenovirus E1A gene transcripts from the Agilent One Color Spike-in kit were added to the isolated RNA (1:50,000). These transcripts, 10 in all, vary 6 logs in concentrations, in one log or half log increments and anneal to complementary control probes on Agilent oligoarrays not present in vertebrate genomes. This served as a quality control metric of the amplification, and subsequent labeling and hybridization protocols. The quality and concentration (typically 7-10 ug) of the amplified single-stranded antisense cDNA products were confirmed and determined by Bioanalyzer Nano Chip electrophoretic traces and on a Nanodrop 2000 spectrophotmeter. Low-quality amplifications were omitted from further analysis.

Oligo microarray and hybridization. We made a custom-designed Agilent zebra finch oligonucleotide microarray (AMDID 022706). The oligo microarray was designed using transcript sequences from a hierarchically organized transcriptome database (http://www.songbirdtranscriptome.net) containing sequences of 91,586 clones isolated from the zebra finch brain[4-6] and several hundred additional transcripts in the NCBI database from various avian species, including known positive and negative control genes regulated by neural activity and targets of FoxP2[72]. The 91,586 clones were clustered into 54,229 relatively unique transcript sequences in the database, including splice variants. For transcripts containing multiple clones, sequences were chosen for probe design based on the quality (average PHRED score of 15 or more) and read length (550 bases or more). Clones with just 3' reads were chosen over clones with just 5' reads, because the 3' end is often preferentially detected in microarray hybridizations. We then filtered out transcripts containing short reads (>150bps) and put a limit of 5 unique transcript variants per gene. This reduced the number of transcripts from 54,229 to 42,304; we then added oligos to other transcripts from NCBI, to obtain a total of 43,552 avian specific transcripts sequences. Probes (60-mers) were then designed to these sequences using Agilent's e-array v5.4 probe selection algorithm. An additional 1,414 control oligos or spots were included (513 Agilent spike in controls described above, 166 triple hairpins that don't binding anything, and 618 dark and bright corner spots, and 120 blank spots placed at strategic locations on the array to normalize occasional hybridization artifacts). This resulted in a total of 44,969 oligo probes/spots on the array. A subset (31,354) of the 43,552 avian oligos matched 9,745 (54%) of the 17,972 predicted gene models by Ensembl (version 54). The 3.2-fold redundancy within the 31,354 transcripts is due to either different oligos generated against the same transcript not detected by EST clustering and to different oligos generated for mRNA variants of the same gene. The remaining 12,484 transcripts on the microarray do not map to Ensembl predicted genes. For microarray analyses, the 43,552 transcripts were re-annotated in a hierarchical manner with the following priority order: 1) Ensembl 54 gene models, 2), manual curations from Wada et al.[6], 3) chicken transcripts in the NCBI database, and 4) human transcripts in the NCBI database.

The arrays were hybridized to linearly amplified cDNAs from song nuclei of singing and silent birds. For all linearly amplified samples, 2.2 ug of cDNA was labeled with Cy3 and purified using the FL-Ovation Cy3 labeling and fragmentation kit. The complementary strand to the spike in controls was added to each sample in the Cy3 labeling reaction. After calculation of the degree of labeling (~3%,) the Cy3 labeled cDNA was fragmented and hybridized overnight using Agilent specified protocols. After hybridization, the slides were washed twice, air dried, and scanned at 5 micron resolution, 532PMT=520, using the Agilent Feature Extraction Software v9.5.1 on a GenePix 4000B microarray scanner (Axon instruments).

Normalization, expression and clustering analysis. The quality of each oligo microarray experiment was determined by analyses of spike-in controls (degree of positive correlation of signal with spike in oligo concentrations) and of probe distribution using the quality assessment program in bioconductor array quality metrics package[73]. Outlier hybridizations that differed significantly from the mean signal due to poor hybridizations or other mechanical factors were removed from further analysis. The number of replicates per time point that survived quality control were 0 (n=6), 0.5hr (n=5) 1hr (n=5), 2hr (n=5), 3hr (n=5), 4hr (n=5), 5hr (n=6), 6hr (n=4), and 7hr (n=5) for a total of 46 microarray experiments after outliers were removed. The data were normalized with variance stabilization[74], which is amenable to the analysis of one-color data. To reduce the noise, probes that were not significantly expressed above two standard deviations of the negative control levels in at least 60% of the arrays were removed from the analysis, leaving 30,119 probes. In cases were there were multiple probes annotated to measure the same gene (splice variants or alternative probes for the same transcript), we merged these probes together if they exhibited the same behavior across measured samples (correlation >= 0.7) through average linkage hierarchical clustering. This merging allowed redundant information to be removed while preserving unique behavior from alternative transcripts, and resulted in 26,071 uniquely behaving transcripts.

To select for genes with significantly regulated gene expression profiles different from null model of no regulation across time, a linear model was used[75]. For each gene, we defined the normalized expression of gene X in array *n* as $X_n = M + X_tT_t + X_bB_n + X_aA_n + e_n$, where M is the mean expression of the gene across all arrays, X_tT_t are time factors, X_bB_n are song behavior factors, X_aA_n are technical RNA amplification factors, and e_n is the residual error of the given array. More specifically, X_t is the time point specific gene expression, and T_t is an indicator variable which is 1 if array n was collected at time t; X_b is the bout dependent gene expression levels and B_n is the number of bouts the bird produced in last half hour before sacrifice, as expression of singing-induced genes is known to depend on the number of bout the bird produces within a specific time[76]; X_a represents how much the measured gene expression level is dependent on amplification error and A_n is an amplification error-dependent term for the specific array (measured by the array quality metric package), to correct for the errors associated with amplifying small amounts of material. This linear model allowed us to use a moderated t-statistic to associate genes with either singing-time or bout number. The p-value was adjusted for multiple hypotheses testing using the false discovery rate. Because each bird sings a different number of bouts, the different samples collected at each time point are not true replicates and were not treated such.

To determine the variety of gene expression patterns, the data were k-means clustered using k=20. We empirically determined that 20 clusters was at or near the optimal, as smaller (5, 10, and 15) and larger (30) clusters did not yield many promoter motifs in the motif search analyses (described below, and not shown).

Cis-regulatory analysis. Of the 807 regulated transcripts in Area X, 439 mapped to ENSEMBLE predicted gene IDs. Because we predicted transcription factor motifs for ENSEMBLE modeled genes

(Supplementary Notes 1), we performed our motif analyses on these 439. From these 439, each cluster (a set of genes with similar expression profiles) was tested for statistical association with cis-regulatory motifs (position weight matrices) and subjected to five statistical control measures (see below). First, we scanned for 99 vertebrate-related motifs from JASPAR[77]. For each motif, the genome was scanned in 500 bp overlapping windows (shifts of 250 bp). Each window was given a HMM-based score for motif clustering[19], as well as a traditional score that counts strong matches to the motif based on log likelihood ratio (LLR). The top 1% of the scoring windows in the genome was considered as motif "target windows". These target windows were then assigned to genes (called the motif's target genes) based on either of two criteria: (a) the window lies within 5 kb upstream or 2 kb downstream of the annotated start site of a gene model, and (b) the window lies in the broader "territory" of a gene defined elsewhere (Supplementary Notes 1). Thus, there are four ways to define a motif's target gene set: two options for window scoring scheme (HMM and LLR), and two options for assigning target windows to genes ('a' and 'b' above). We traversed the sorted list of target windows, assigning windows to their respective genes and designating them as motif targets, until 500 distinct genes had been designated as the motif's target genes. The universe of genes considered was all 9,745 genes represented on the microarrays described above. For each of the four ways of defining target genes, the motif's target set was tested for enrichment in the gene set defined by a cluster, using a hypergeometric test. Thus, for each cluster, 4 motif gene sets x 99 motifs = 396 statistical tests were performed for associations with motifs, and equally many p-values were obtained (statistical test 1 below). These were then assigned q-values[78], to correct for multiple hypotheses testing, using the R package (statistical test 2 below). In addition, two types of negative controls were performed, one where the motif target set was replaced in a simulation experiment with random gene sets of similar size and characteristics (statistical test 3 below), and another where the cluster was replaced with random sets of genes that are not regulated by singing activity (statistical test 4 below). Based on these four statistical test, motif–cluster associations were chosen as significant.

In a separate analysis, 18 activity-dependent motifs from JASPAR and TRANSFAC[79] databases (listed below; + from TRANSFAC) were selected based on prior knowledge about the neural activity-dependence of their corresponding transcription factors[80]. The 18 motifs belong to six different activity-dependent transcription factor families.

Transcription factor family		**Motif**
1. cAMP Response Element Binding Protein	1	CREB1
	2	+CREBP1
	3	+CREB-ATF
	4	+CREBP1-CJUN
2. Serum Response Factor	5	SRF
3. Activity-dependent Transcription Factors	6	+ATF
	7	+ATF1
	8	+ATF4
4. Myocyte Enhancer Factors	9	MEF2A

	10	+AMEF2
	11	+HMEF2
	12	+MEF2
5. Nuclear Factor Kappa B	13	NFKB
	14	NFKB1
6. Immediate early gene transcription factors	15	Fos
	16	+AP1
	17	+AP1(fos-jun)
	18	+EGR1

The first five families are transcription factors post-translationally modified by neural activity. The sixth family, the immediate early gene transcription factors, is transcriptionally modified by neural activity. Associations were determined between these motifs and each cluster, using the same four statistical tests as mentioned above. Details of the four tests are as follows:

1. **Hypergeometric test p-value.** The key statistical test performed was the Hypergeometric test of intersection between a given cluster and the "target gene set" of a particular motif M. A significant p-value suggest an association between the cluster and the motif ("hg_p-Value" of Supplementary Table 6)

2. **Hypergeometric test q-value**. For any given cluster, analysis (1) involved computing p-values for several motifs and four different methods for each motif, thus leading to a need for multiple hypothesis correction. Such correction was performed using Storey's "q-value" calculations ("hg_q-Value" of Supplementary Table 6).

3. **Simulation Hypergeomtric p-value**. We computed a "simulation-based p-value" for each cluster-motif pair. In the Hypergeometric test for a fixed cluster of genes, we examined its overlap (say K genes) with a motif's target gene set (say N genes) and asked "What is the probability that a random selection of N genes would have an overlap with the cluster that is ≥ K"? Here, we answer the same question through simulations: i.e., randomly sample a "target gene set" (again of N genes) for the motif, note the overlap with the cluster, repeat 1000 times, and report the fraction of times that the overlap was found to be ≥ K. This results in a simulation-based p-value (Supplementary Table 6).

4. **Non-regulated random cluster control:** We performed a control experiment where the singing regulated gene clusters were replaced with random sets of genes of similar cluster size, and ensured that these random clusters do not show an association with singing (although an individual singing regulated gene could be randomly pick as part of a cluster). When studying any particular singing-dependent cluster, **C**, suppose we obtain a low p-value p_M for a particular motif

M. We tested whether this low p-value would not have been seen when we replace the cluster **C** with a random, non-regulated set of genes. To perform this analyses, we (a) choose a set of genes **C'** (of same cardinality as **C**) from all genes on the array that are not in any of the 20 clusters, i.e., which are not singing regulated; (b) test this cluster **C'** against all motifs (e.g. all 99 JASPAR motifs) and all four methods, obtaining a large number of p-values corresponding to **C'**; (c) count how many p-values in this list are below p_M; (d) repeat steps (a) to (c) 100 times, i.e., for 100 random clusters, and compute the average number of p-values better than p_M; and (e) use this average to obtain an empirical estimate of the "false discovery rate" corresponding to p-value p_M (est-FDR column, Supplementary Table 6).

With the above four tests, we then chose significant motifs. For the analyses with the 99 JASPAR motifs, we required an uncorrected hg p-value < 0.02, a sim p-value < 0.02, and either a hg q-value < 0.25 or an est-FDR < 0.25. The motifs that met these criteria were considered to be high quality motifs. For the analyses with 18 activity dependent motifs, all multiple hypothesis corrections were performed based on the 4 test x 18 motifs = 72 tests (p-values) for each cluster. Thus, we increased the stringency of our significance criteria for these high-quality motifs and required an hg p-value < 0.01, sim p-value > 0.01, a hg q-value <0.25 and est-FDR < 0.25.

Shuffled activity-dependent motif control: In addition to the above four tests, we performed an additional negative control analysis with the known 18 activity-dependent motifs. We replaced the 18 motifs of interest with "randomly shuffled" versions that represent entirely different (randomly chosen) binding specificities. This resulted in a new compendium of 18 motifs to mirror the original analysis, except that the motifs in this compendium are not "real". We took care that the *information content* of each shuffled motif matches that of an original motif, i.e., the shuffled motif compendium has the same rough characteristics as the original compendium. We repeated the entire analysis, and count, both for the original analysis and the new (control) analysis, the number of motifs at various levels of significance.

Based on the above analyses, we report the following additional results:

Dynamic cascades of singing-behavior driven gene expression. As described in Figure 5, we identified a set of 807 genes with significantly altered singing-time series expression dynamics (FDR < 0.05). This included several genes whose expression dynamics matched their known time points of singing-induced expression in prior *in-situ* hybridization studies[76], including *egr1* (a.ka. *ZENK), c-fos*, and *Arc*, peaking at 0.5hrs, and *FoxP2* decreasing at 2 hrs. We clustered this set into 20 groups with k-means clustering. Examining the 20 clusters of all significantly expressed gene profiles reveals a pattern of expression that includes, rapidly up-regulated and transient over time (cluster 1 peaking at 0.5hr), rapidly up-regulated and sustained over time (clusters 2-4 on at 0.5 hr), slowly up-regulated and sustained over time (clusters 5-8 onset between 1-3 hrs), and a late response up-regulated genes (clusters 9-12 on at 3 hr; Figure 5a and Supplementary Table 6). We also found an equally diverse set of down-regulated clusters of genes (clusters 13-20; Figure 5a and Supplementary Table 6).

We found significant over-representation of specific promoter motifs in 14 of the 20 gene clusters (Figure 5a; Supplementary Table 6, using both the 99 JASPAR and 18 selected activity-dependent motifs). The rapid, transiently upregulated genes of cluster 1 had an over-representation of the cis-regulatory binding site for the post-translationally activity-dependent transcription factor combination *CREB-ATF*, and others not known or yet studied for their activity-dependence. In turn, the binding sites for the immediate early gene transcription factors in the early clusters (such as cluster 1 for fos, and ATF1, ATF4, and c-jun known to be regulated by singing) were over-represented in late response gene clusters 9-12 (e.g., Cluster 10 for *c-fos* and *c-jun* as *AP1* [*fos-jun* dimer]; Supplementary Table 6). Surprisingly also over-represented in the cis-regulatory regions of late response up-regulated clusters were binding motifs for multiple post-translationally activated activity-dependent transcription factors (e.g., *CREBP1*, *NFKB1, AMEF2* and *SRF*) which have been shown to activate different immediate early genes, but in cells in culture[80]. Even more remarkable was that the late response up-regulated clusters were extremely enriched for the post-translational and transcriptionally activated activity-dependent transcription factors relative to all other gene clusters (Figure 5a; yellow shading in Supplementary Table 6). These results show the potential strength of our motif association analysis, and that the late wave activity-dependent gene expression may not simply be the result of activation by induced immediate early gene transcription factors, but a combination of post-translationally and transcriptionally activated transcription factors.

Compared to the up-regulated gene clusters 1-12, where 26 over-represented motifs in total were found, only 7 motifs were found as over-represented in the clusters 13-20 of down regulated genes, and only one, Foxq1, is known to be activity regulated (Supplementary Table 6). These results suggest that there may be undiscovered transcription factor motifs for transcriptional suppression or a transcription factor independent mechanism for down-regulated genes.

When we randomly shuffled the 18 activity-dependent motifs followed by genome-wide scanning for targets of these random motifs, we nearly eliminated finding significant associations. In particular, with the 18 real motifs, there were 14 motif-cluster associations that passed the criterion of our four statistical tests (Supplementary Table 6), but only one association in the shuffled analysis (CREBP1 for random cluster 16). These findings demonstrate that the motifs discovered represent potential real singing regulated genome targets and should allow future investigations on reconstructing behavior-dependent gene regulatory networks.

All gene expression analyses are found at the GEO database accession number GSE19900. Further analyses is available here: http://aviangenomes.org

Supplementary Note 4. Adaptive evolution of vocal communication genes

In an analysis of the Ensembl 54 zebra finch gene set, a branch-site model implemented in PAML[81] identified 1,168 genes containing specific residues under positive selection (PS) in the zebra finch lineage. 632 of these PS genes are represented in the microarray data of Dong et al.[68]. 410 are significantly reduced in expression (FDR= 5%) in the auditory forebrain after novel song exposure, compared to birds hearing either silence or familiar song, and are here termed song suppressible (SS) genes. The intersection of these two lists (PS+SS) yields 49 genes (Supplementary Table 7).

An analysis was performed to assess the composition of this PS+SS gene set using the Gene Ontology (GO) annotations in Ensembl 54 and CORNA software[82], as implemented in a public software tool (http://*Bioinformatics*.iah.ac.uk/tools/Gofinch). The reference population was the 4,640 genes that were both represented on the array by at least one EST that overlaps the gene model as described in S3 Supplementary notes, and assessed for positive selection in the initial PAML ortholog analysis. The results are given here in Supplementary Table 8, showing all the terms that are significant at adjusted $p < 0.05$, and Supplementary Table 9, showing the genes associated with those terms. Note that eleven genes contribute redundantly to these top terms. Six are annotated for ion channel activity as shown in Table 2. Two others have annotations suggesting a function in ion channel biology although they do not currently have “ion channel activity” as a GO annotation term: *TRPM7-2* (Transient receptor potential cation channel subfamily M member 7) and *ZFHX4* (Zinc finger homeobox protein 4).

Considering just the six genes explicitly annotated for ion channel activity (GO:0005216), the p-value for finding 6 ion channel genes in a random sample of 49 is $1.5\text{x}10^{-05}$ or 0.0016 with adjustment for multiple testing, given the total number of genes (49) on the microarray sharing this annotation term (Supplementary Table 8). We also considered the significance of the interaction between the two identification criteria, PS and SS. That is, once one criterion had been applied to reduce the reference population, what were the chances that applying the other criterion would result in inclusion of 6 ion channel genes on the list of 49? Using just the 632 PS genes on the array as the reference population, the p-value that a sample of 49 will then contain 6 ion channel genes is $1.5\text{x}10^{-04}$ (0.026 adjusted). Using just the 410 SS genes as the reference population, the probability of selecting 6 ion channel genes in a sample of 49 is 0.036 (0.026 adjusted). We conclude that there is a significant interaction across all three conditions: positively selected residues in the finch lineage; suppressed by song in the auditory forebrain; and encoding a protein with ion channel activity.

We performed equivalent analyses for overlap between PS genes and the other gene regulation sets defined in Dong et al.[68]. In total, 214 PS genes on the microarray are also responsive in some way to song. Taking this set as a whole, no GO term is significant at adjusted $p < 0.05$ (89 GO terms are significant at an FDR of 20%). Nor is any GO term significant at adjusted $p < 0.05$ for any of the other gene regulation subsets.

We performed additional dN/dS analyses for the genes shown in Table 2 and Supplementary Table 9, using PAML[32,83] to analyze alignments of Ensembl gene models from seven species: zebra finch, chicken, anolis, human, mouse, opossum, and rat. We first tested for evidence of variation in ω (or the ratio of nonsynonymous to synonymous substitutions or dN/dS) using ‘branch’ models, which test for selection acting across the entire gene. We compared a model with two rates, in which birds and non-birds were each allowed to have a unique rate, to a model with three rates, in which the zebra finch also was also allowed to have a unique rate. By comparing these models using a likelihood ratio test (LRT), we identified genes that were evolving at a unique rate in zebra finch. Because selection often acts on the level of specific codons, rather than across whole genes, we also examined ion channel genes using a branch-sites model[84]. We used an LRT to compare a model in which sites in the zebra finch lineage were allowed $\omega > 1$ with a null model in which ω was constrained to ≤ 1. Empirical Bayes analysis[84] was used to estimate the posterior probability that specific sites within each gene have $\omega > 1$ in the zebra finch lineage. We also ran both branch and branch-sites analyses using manually curated gene models. These analyses were broadly consistent with analyses using Ensembl models.

II. Supplementary Tables

Supplementary Table 1. Assembly metrics of contiguity

Genome Feature	Chicken	Zebra finch
N50 contig length	36kb	39kb
N50 supercontig length	7Mb	10Mb
EST coverage [a]	96/85%	96/91%
Assembled bases	1.06Gb	1.2Gb
Bases localized to chromosome	933Mb	1.0Gb

[a] At least partially aligned/>=50% aligned

Supplementary Table 2. Synteny analysis of Syn1.

Human (X-chromosome)	Zebra finch	Chicken	Lizard
FUNDC1 (44.27 Mb)	FUNDC1 (1: 5.6 Mb)	FUNDC1 (1: 114.59 Mb)	FUNDC1 (Scaffold 118)
DUSP21 (44.59 Mb)			……
UTX (44.62 Mb)	UTX1 (1: 5.4 Mb)	UXT1 (1: 114.40 Mb)	UTX (Scaffold 118)
ENSG00000215290 (44.68 Mb)	No homologues	No homologues	……
Q8WZ11_HUMAN (44.89 Mb)	No homologues	No homologues	……
CXorf36 (44.90 Mb)	CXorf36 (1: 5.4 Mb)	CXorf36 (1: 114.37 Mb)	CXorf36 (Scaffold 118)
ENSG00000215287 (45.38 Mb)	No homologues	No homologues	……
ZNF673 (46.19 Mb)	No homologues	No homologues	……
ZNF674 (46.24 Mb)	No homologues	No homologues	……
CHST7 (46.32 Mb)	No homologues	No homologues	……
ENSG00000204910 (46.35 Mb)	No homologues	No homologues	……
SLC9A7 (46.35 Mb)	SCL9A7 (1: 33.49 Mb)	SLC9A7 (1: 134.38 Mb)	SLC9A7 (Scaffold 571)
RP2 (46.58 Mb)	RP2 (1: 33.58 Mb)	XRP2_CHICK (1: 134.31 Mb)	RP2 (Scaffold 571)
CXorf31 (46.63 Mb)	……	……	……
PHF16 (46.66 Mb)	PHF16 (1: 33.64 Mb)	PHF16 (1: 134.25 Mb)	PHF16 (Scaffold 571)
RGN (46.82 Mb)	RGN (1: 33.68 Mb)	RGN_CHICK (1: 134.23 Mb)	RGN (Scaffold 571)
NDUFB11 (46.89 Mb)	*	No homologues	NDUFB11 (Scaffold 158)
RBM10 (46.89 Mb)	No homologues	No homologues	RBM10 (Scaffold 158)
ENSG00000215281 (46.93 Mb)	No homologues	No homologues	……
UBE1 (46.94 Mb)	No homologues	No homologues	UBE1 (Scaffold 158)

INE1 (46.95 Mb)	No homologues	No homologues	……
PCTK1 (46.96 Mb)	No homologues	No homologues	PCTK1 (Scaffold 158)
USP11 (46.98 Mb)	No homologues	No homologues	USP11 (Scaffold 158)
ZNF157 (47.11 Mb)	No homologues	No homologues	……
NPM1 (47.18 Mb)	NPM1 (13: 1.5 Mb)	NPM_CHICK (13: 3.03 Mb)	NPM1 (Scaffold 109)
ZNF41 (47.19 Mb)	No homologues	No homologues	……
CXorf24 (47.23 Mb)	No homologues	No homologues	……
ARAF (47.31 Mb)	No homologues	No homologues	ARAF (Scaffold 248)
Timp1 (47.44 Mb)	No homologues	No homologues	TIMP1 (Scaffold 248)
SYN1 (47.31 Mb)	No homologues	No homologues	SYN1 (Scaffold 248)
CFP (47.48 Mb)	No homologues	*	CFP (Scaffold 248)
ELK1 (47.49 Mb)	No homologues	No homologues	ELK1 (Scaffold 158)
UXT (47.51 Mb)	*	No homologues	UXT (Scaffold 158)
CXorf25-201 (47.58 Mb)	No homologues	No homologues	……
ALO22578	No homologues	No homologues	……
ZNF81	No homologues	No homologues	……
ZNF182	No homologues	No homologues	……
ZNF630	No homologues	No homologues	……
SPACA5 (47.98 Mb)	No homologues	No homologues	……
SSX3 (48.20 Mb)	No homologues	No homologues	……
SSX4 (48.24 Mb)	No homologues	No homologues	……
SSX4B (48.26 Mb)	No homologues	No homologues	……
SLC38A5 (48.31 Mb)	SCL38A5 (1: 30 Mb)	SCL38A5 (Un_random)	SLC38A5 (Scaffold 158)

Syn1 and its flanking syntenic group are missing in birds. The region of the human X chromosome containing *Syn1* and flanking genes (left column) is shown in comparison with birds (chicken and zebra finch; middle columns) and lizard (green anole; right column). Genes in light green could be found and constitute a syntenic group in all four species. In contrast, genes in yellow, including *Syn1* (in orange) and its immediately flanking genes, could be found only in the genomes of humans (and other mammals) and lizard (as well as in frogs and fish) but not in birds. Note that *NPM1* (in light blue) is present in all four species but it is not syntenic with any other genes on this table (except for humans and other mammals), indicating that its proximity to *Syn1* is specific to the mammalian lineage. The remaining genes on this table (no shading) could only be found in mammalian genomes. We note that we have identified avian ESTs (from zebra finch, chicken and/or turkey) that appear to be related to the genes indicated with an asterisk (*NDUFB11, CFP* and *UXT*). However, the overall identities of these transcripts compared to the predicted human proteins are rather low (~45%). In addition, these ESTs did not align to the current avian genomic sequences nor had any significant hits to avian genomic trace archives; it is thus presently unclear whether these ESTs represent true avian orthologs for these genes. We did not obtain avian EST evidence for any other genes within the region between *NDUFB11* and *UXT* on the human X. Finally, extensive searches of zebra finch brain cDNA libraries have failed to generate any evidence of a brain transcript representing an avian ortholog of *Syn1*, which is an abundantly expressed gene in the brain of mammals.

Supplementary Table 3. A summary of the zebra finch degradome.

Gene	Process	Zebra finch	Chicken	Human
Apoptosis				
Caspase-1, -4, -5, -12	Apoptosis and inflammation	Not found	Only caspase-1	All
Caspase-17		Not found	Present	Absent
Caspase-18		Present	Present	Absent
Host defense				
Granzyme K	Cytotoxic protease	Absent	Absent	Present
Neutrophil elastase	Neutrophil antibiotic protein	Absent	Absent	Present
Complement factor D		Absent	Absent	Present
Azurocidin	Neutrophil antibiotic protein	Absent	Absent	Present
Proteinase 3	Neutrophil antibiotic protein	Absent	Absent	Present
Granzymes B and H, cathepsin G, chymase	Cytotoxic protease, others	Granzyme Z and Z-like	Granzyme Z	All
EOS		Absent	Absent	Present
Tryptases	Mast cell biology	Absent	Absent	Present
Haptoglobins		Absent	Absent	Present
Cathepsin F		Absent	Absent	Present
Cathepsin W	Regulation of T-cell cytolytic activity	Absent	Absent	Present
PRSS16	Alternative antigen presenting pathway in thymus	Absent	Absent	Present
Paracaspase	NFkB activation	Three copies	Three copies	Present
Legumain-2	MHC class II presentation	Absent	Absent	Present
Tissue development				
Enamelysin	Enamel formation	Absent	Absent	Present
Kallikrein-4	Enamel formation	Absent	Absent	Present
MMP-7, -8, -19, -21, -23B, -25, -26	Extracellular matrix remodeling	Absent	Absent	Present
ADAMTS-4	Aggrecan degradation	Absent	Absent	Present
Reproduction				
Alpha-aspartyl dipeptidase	Reproduction	Present	Present	Absent
Nothepsin	Egg development	Present	Present	Absent
ADAMTS-16		Absent	Absent	Present
Testins	Zona pellucida lysis	Absent	Absent	Present
ADAM3B, -4,-4B, -5, -6, -7	Sperm-egg interaction	Absent	Absent	Present
ADAM30	Sperm-egg interaction	Absent	Absent	Present
Acrosin	Sperm-egg interaction	Expanded (7 genes)	Present	Present
Prolactin-induced protein	Pregnancy	Absent	Absent	Present
Implantation serine proteases	Embryo implantation	Absent	Absent	Absent (present in mouse)
Ovastacin	Embryo hatching	Absent	Absent	Present
Choryolytic enzymes	Embryo hatching	Present	Present	Absent
Neural development				
Caspase-3	Learning and memory	Duplicated	Present	Present
Neurotrypsin	Learning and memory	Duplicated	Present	Present
Bace	Cleavage of amyloid precursor protein	Duplicated	Present	Present
Presenilin homolog-2	Cleavage of amyloid precursor protein	Absent	Absent	Present
Transmembrane serine protease 5		Absent	Present	Present
Brain serine protease-2		Absent	Absent	Present
Other				
Pepsinogen A	Protein digestion in the stomach	Duplicated	Duplicated	Triplicated
Pepsinogen C	Protein digestion in the stomach	Not found	Present	Present
ADAMTS-13	Cleavage of von Willebrand Factor	Duplicated	Present	Present
Tubulointerstitial nephritis antigen-like 1	Renal function	Absent	Absent	Present
Desert hedgehog protein	Morphogenesis	Absent	Absent	Present
Sentrin-3	Desumoylation	Absent	Absent	Present
Autophagin-4	Autophagy	Absent	Absent	Present
NAALADASE like 1		Absent	Absent	Present
Transferrin receptor 2	Cellular uptake of transferrin-bound iron	Absent	Absent	Present (homolog)
dipeptidyl-peptidase 3	Ovarian function	Absent	Absent	Present
Vitellogenic-like carboxypeptidase		Absent	Absent	Present
ClpP caseinolytic peptidase		Absent	Absent	Present
Abhydrolase domain containing 4		Absent	Absent	Present
Polyserases		Only polyserase-1	Only polyserase-1	Polyserase-1, -2, -3
Kallikreins (15 genes)		Absent	Absent	Present

Supplementary Table 4. RepeatMasker analysis for chicken and zebra finch.

	taeGut1 (1.049 maskable Gb)			galGal3 (1.044 maskable Gb)		
	Copies	Kb	Fraction	Copies	Kb	Fraction
LINE/CR1	133,331	39,174	3.73%	187,463	65,448	6.27%
SINE/MIR	2,721	287	0.03%	2,659	279	0.03%
SINE/tRNA-CR1	2,791	280	0.03%	0	0	0.00%
LTR/ERV1	12,962	6,759	0.64%	2,229	1,481	0.14%
LTR/ERVK	36,629	19,948	1.90%	1,119	819	0.08%
LTR/ERVL	28,923	14,530	1.38%	26,689	12,416	1.19%
DNA transposons	186	56	0.01%	16,624	8,446	0.81%
Total IRs	217,543	81,035	7.72%	236,783	88,890	8.52%
Satellite	488	85	0.01%	4,426	2,293	0.22%
Simple repeats	123,377	6,829	0.65%	137,348	6,394	0.61%
Low complexity	196,866	8,734	0.83%	139,270	5,734	0.55%

Supplementary Table 5. **Estimation of repeat-derived fraction and the ERV1-derived fraction of the zebra finch using *P-Clouds***.

Analysis[a]	Mbp[b]	False Positive[c]	Sensitivity[d]	Adjusted Mbp[e]	Repetitive Proportion[e,f]
General P-Clouds	439.87	2.68%	78.7%	544.0	51.86%
ERV1-Specific P-Clouds	2.67	4.5%	90.6%	10.13	0.96%[g]

[a] Parameter settings were (2,8,16,160,1600) for general *P-Clouds* and (2,2,4,40,400) for ERV1-specific *P-Clouds*.

[b] For ERV1-specific *P-Clouds*, this is the Mbp found in the region not annotated by *RepeatMasker*.

[c] All false positive rates are estimated from re-synthesis of the region not annotated by *RepeatMasker*.

[d] Estimated as the proportion of *RepeatMasker*-identified bp not annotated.

[e] All estimated repetitive Mbp and proportions are adjusted based on the false positive rate; for the ERV1-specific *P-Clouds*, it is also adjusted based on the sensitivity, and the total genomic ERV1 region is given.

[f] All proportions assume a maskable genome size of 1.049 Mbp.

[g] This is a 39% increase compared to the *RepeatMasker* estimate alone.

Supplementary Table 6. Transcription factor binding sites enriched in clusters of singing regulated genes in Area X.

expression pattern	#	motifs	hg p-value	sim p-value	hg q-value	est-FDR
Rapidly up-regulated & transient	1	**+CREB-ATF**	**0.00244**	**0.0012**	**0.176**	**0.02**
		IRF1	0.00244	0.0047	0.274	0.08
		MafB	0.00244	0.0047	0.274	0.04
		Myf	0.00244	0.0047	0.274	0.026
Rapidly up-regulated & sustained	2	Sox17	0.00108	0.0013	0.427	0.22
	3	Myf	0.00068	0.0007	0.135	0.27
		Roaz	0.00068	0.0021	0.135	0.54
		REL	0.0033	0.0039	0.187	0.73
		SPI1	0.0033	0.0039	0.187	0.5475
		TAL1-TCF3	0.0033	0.0089	0.187	0.3186
		Foxd3	0.0033	0.0089	0.187	0.43
Slowly up-regulated & sustained	4	none				
	5	TCF1	0.00006	0.0001	0.022	0.01
	6	none				
	7	none				
	8	Nkx2-5	0.00142	0.0053	0.282	0.18
Late response up-regulated	9	**+CREBP1**	**0.00698**	**0.0055**	**0.168**	**0.08**
		+CREBP1-CJUN	**0.00698**	**0.0055**	**0.168**	**0.12**
	10	**+NFKB1**	**0.00019**	**0.0002**	**0.014**	**0.001**
		+CREBP1	**0.00788**	**0.0061**	**0.142**	**0.16**
		+AP1(cfos-cjun)	**0.00788**	**0.0061**	**0.142**	**0.21**
	11	**+SRF**	**0.00002**	**0.0003**	**0.001**	**0.001**
		+AMEF2	**0.00019**	**0.0015**	**0.007**	**0.015**
		+ATF1	**.00171**	**0.0089**	**0.037**	**0.053**
		Hox11-CTF1	0.0171	0.008	0.226	0.226
	12	MYC-**MAX**	**0.00171**	**0.0031**	**0.339**	**0.13**
		+ATF4	**0.00171**	**0.0031**	**0.062**	**0.005**
		+SRF	**0.00171**	**0.0031**	**0.062**	**0.01**
		+CREB-ATF	**0.01179**	**0.0187**	**0.170**	**0.046**
		+ATF	**0.01179**	**0.0187**	**0.170**	**0.046**
Mixed up- & down-regulated	13	RORA-2	0.00129	0.0026	0.170	0.04
		MIZF	0.00129	0.0031	0.170	0.08
		RXR-VDR	0.00129	0.0031	0.170	0.027
Rapidly down-regulated & transient	14	ZEB1	0.00124	0.001	0.492	0.01
	15	Lhx3	0.0002	0.0017	0.081	0.01
Rapidly down-regulated & sustained	16	none				
Slowly down-regulated & sustained	17	TFAP2A	0.00081	0.0006	0.323	0.25
	18	none				
	19	**Foxq1**	**0.00025**	**0.0009**	**0.100**	**0.38**
Late response down-regulated	20	none				

Clusters are ordered according to temporal expression profile. Light red to dark red shadding, multiple clusters of up-regulated genes. Light blue to dark blue shadding, multiple clusters of down regulated genes. Alternating grey bars, the 20 k-means clusters of singing regulated genes. Shown are the motifs each cluster that passed our significance criterion (hg and sim p-values < 0.02, and either a hg q-value < 0.25 or an est-FDR < 0.25 for the 99 JASPAR motifs; and hg and sim p-value < 0.01for the 18 activity-dependent motifs). Yellow shading, known activity-regulated transcription factors are highlighted; bold brown text, motifs of transcription factors known to be post-translationally modified by neural activity that then induce target genes, including those found as regulated by singing; bold black text, the target transcription factors whose mRNA expression is up- or down-regulated by neural activity, including by singing in Area X as found in the 807 genes or previous reports. Some transcription factors co-regulate the same motifs in dimer combinations (e.g. *AP1FJ* for *fos* and *jun*; and *CREB-ATF*). +, motifs from the TRANSFAC data base; all others are from JASPAR.

Supplementary Table 7. The 49 genes that are both suppressed in expression by song stimulation in the analysis of Dong et al. [68], and show evidence of positive selection in the zebra finch lineage in the analysis of [85].

Ensembl Gene ID	Associated Gene Name	Description
ENSTGUG00000002344	ABCA2	ATP-binding cassette sub-family A member 2 (ATP-binding cassette transporter 2)(ATP-binding cassette 2)
ENSTGUG00000001167	ARHGAP21	Rho GTPase-activating protein 21 (Rho-type GTPase-activating protein 21)(Rho GTPase-activating protein 10)
ENSTGUG00000003548	BICD2	Protein bicaudal D homolog 2 (Bic-D 2)
ENSTGUG00000011342	C2orf21	Protein unc-80 homolog [Source:UniProtKB/Swiss-Prot;Acc:Q8N2C7]
ENSTGUG00000002855	CACNA1B	Voltage-dependent N-type calcium channel subunit alpha-1B (Voltage-gated calcium channel subunit alpha Cav2.2)(Calcium channel, L type, alpha-1 polypeptide isoform 5)(Brain calcium channel III)(BIII)
ENSTGUG00000009049	CACNA1G	Voltage-dependent T-type calcium channel subunit alpha-1G (Voltage-gated calcium channel subunit alpha Cav3.1)(Cav3.1c)(NBR13)
ENSTGUG00000008314	CCDC41	Coiled-coil domain-containing protein 41 (Renal carcinoma antigen NY-REN-58)
ENSTGUG00000007769	CDKL5	Cyclin-dependent kinase-like 5 (EC 2.7.11.22)(Serine/threonine-protein kinase 9)
ENSTGUG00000001731	CLASP2	CLIP-associating protein 2 (Cytoplasmic linker-associated protein 2)(hOrbit2)
ENSTGUG00000003971	CREBBP	CREB-binding protein (EC 2.3.1.48)

ENSTGUG00000012492	DACH1	Dachshund homolog 1 (Dach1)
ENSTGUG00000010011	DNAJC6	Putative tyrosine-protein phosphatase auxilin (EC 3.1.3.48)(DnaJ homolog subfamily C member 6)
ENSTGUG00000010112	DVL3	Segment polarity protein dishevelled homolog DVL-3 (Dishevelled-3)(DSH homolog 3)
ENSTGUG00000003933	EPAS1	Endothelial PAS domain-containing protein 1 (EPAS-1)(Member of PAS protein 2)(Basic-helix-loop-helix-PAS protein MOP2)(Hypoxia-inducible factor 2 alpha)(HIF-2 alpha)(HIF2 alpha)(HIF-1 alpha-like factor)(HLF)
ENSTGUG00000010817	ERCC5	DNA repair protein complementing XP-G cells (Xeroderma pigmentosum group G-complementing protein)(DNA excision repair protein ERCC-5)
ENSTGUG00000003051	FAM13A1	Protein FAM13A1
ENSTGUG00000011061	FARP1	FERM, RhoGEF and pleckstrin domain-containing protein 1 (Chondrocyte-derived ezrin-like protein)
ENSTGUG00000000634	GLB1L2	Beta-galactosidase-1-like protein 2 Precursor (EC 3.2.1.-)
ENSTGUG00000000694	GPR98	G-protein coupled receptor 98 Precursor (Monogenic audiogenic seizure susceptibility protein 1 homolog)(Very large G-protein coupled receptor 1)(Usher syndrome type-2C protein)
ENSTGUG00000005484	GRIA2	Glutamate receptor 2 Precursor (GluR-2)(GluR-B)(GluR-K2)(Glutamate receptor ionotropic, AMPA 2)(AMPA-selective glutamate receptor 2)
ENSTGUG00000003563	GRIA3	Glutamate receptor 3 Precursor (GluR-3)(GluR-C)(GluR-K3)(Glutamate receptor ionotropic, AMPA 3)(AMPA-selective glutamate receptor 3)

ENSTGUG00000007354	KCNC2	Potassium voltage-gated channel subfamily C member 2 (Voltage-gated potassium channel Kv3.2) [Source:UniProtKB/Swiss-Prot;Acc:Q96PR1]
ENSTGUG00000008921	LRRC16A	Leucine-rich repeat-containing protein 16A (CARMIL homolog)
ENSTGUG00000009531	MAP2K5	Dual specificity mitogen-activated protein kinase kinase 5 (MAP kinase kinase 5)(MAPKK 5)(EC 2.7.12.2)(MAPK/ERK kinase 5)
ENSTGUG00000008126	MOSPD2	Motile sperm domain-containing protein 2
ENSTGUG00000005201	MYH10	Myosin-10 (Myosin heavy chain 10)(Myosin heavy chain, non-muscle IIb)(Non-muscle myosin heavy chain IIb)(NMMHC II-b)(NMMHC-IIB)(Cellular myosin heavy chain, type B)(Non-muscle myosin heavy chain-B)(NMMHC-B)
ENSTGUG00000012213	NEK9	Serine/threonine-protein kinase Nek9 (EC 2.7.11.1)(Never in mitosis A-related kinase 9)(NimA-related protein kinase 9)(Nercc1 kinase)(NIMA-related kinase 8)(Nek8)
ENSTGUG00000007055	NFX1	Transcriptional repressor NF-X1 (EC 6.3.2.-)(Nuclear transcription factor, X box-binding protein 1)
ENSTGUG00000005630	NRCAM	*Neuron*al cell adhesion molecule Precursor (Nr-CAM)(NgCAM-related cell adhesion molecule)(Ng-CAM-related)(hBravo)
ENSTGUG00000009501	PDE1A	Calcium/calmodulin-dependent 3',5'-cyclic nucleotide phosphodiesterase 1A (Cam-PDE 1A)(EC 3.1.4.17)(61 kDa Cam-PDE)(hCam-1)
ENSTGUG00000000146	PIGG	GPI ethanolamine phosphate transferase 2 (EC 2.-.-.-)(Phosphatidylinositol-glycan biosynthesis class G protein)(PIG-G)(GPI7 homolog)(hGPI7)
ENSTGUG00000004740	PTPRD	Receptor-type tyrosine-protein phosphatase delta Precursor (Protein-tyrosine phosphatase delta)(R-PTP-delta)(EC 3.1.3.48)

ENSTGUG00000006417	RBM5	RNA-binding protein 5 (RNA-binding motif protein 5)(Tumor suppressor LUCA15)(Protein G15)(Renal carcinoma antigen NY-REN-9)
ENSTGUG00000011412	SMOC1	SPARC-related modular calcium-binding protein 1 Precursor (Secreted modular calcium-binding protein 1)(SMOC-1)
ENSTGUG00000006871	SORBS2	Sorbin and SH3 domain-containing protein 2 (Arg/Abl-interacting protein 2)(ArgBP2)(Sorbin)
ENSTGUG00000002680	SP4	Transcription factor Sp4 (SPR-1)
ENSTGUG00000008287	SPHKAP	A-kinase anchor protein SPHKAP (SPHK1-interactor and AKAP domain-containing protein)(Sphingosine kinase type 1-interacting protein)
ENSTGUG00000011841	STRN3	Striatin-3 (Cell-cycle autoantigen SG2NA)(S/G2 antigen)
ENSTGUG00000000079	TCERG1	Transcription elongation regulator 1 (TATA box-binding protein-associated factor 2S)(Transcription factor CA150)
ENSTGUG00000004694	TEX2	Testis-expressed sequence 2 protein
ENSTGUG00000010899	TIMP3	Metalloproteinase inhibitor 3 Precursor (Tissue inhibitor of metalloproteinases 3)(TIMP-3)(Protein MIG-5)
ENSTGUG00000003073	TLK2	Serine/threonine-protein kinase tousled-like 2 (EC 2.7.11.1)(Tousled-like kinase 2)(PKU-alpha)
ENSTGUG00000001431	TMEM39B	Transmembrane protein 39B

ENSTGUG00000003997	TP53BP2	Apoptosis-stimulating of p53 protein 2 (Tumor suppressor p53-binding protein 2)(p53-binding protein 2)(p53BP2)(53BP2)(Bcl2-binding protein)(Bbp)(Renal carcinoma antigen NY-REN-51)
ENSTGUG00000007440	TRPM7-2	Transient receptor potential cation channel subfamily M member 7 (EC 2.7.11.1)(Long transient receptor potential channel 7)(LTrpC7)(Channel-kinase 1)
ENSTGUG00000006952	TRPV1	Transient receptor potential cation channel subfamily V member 1 (TrpV1)(Osm-9-like TRP channel 1)(OTRPC1)(Vanilloid receptor 1)(Capsaicin receptor)
ENSTGUG00000011165	UBR1	E3 ubiquitin-protein ligase UBR1 (EC 6.3.2.-)(N-recognin-1)(Ubiquitin-protein ligase E3-alpha-1)(Ubiquitin-protein ligase E3-alpha-I)
ENSTGUG00000005529	YEATS2	YEATS domain-containing protein 2
ENSTGUG00000011605	ZFHX4	Zinc finger homeobox protein 4 (Zinc finger homeodomain protein 4)(ZFH-4)

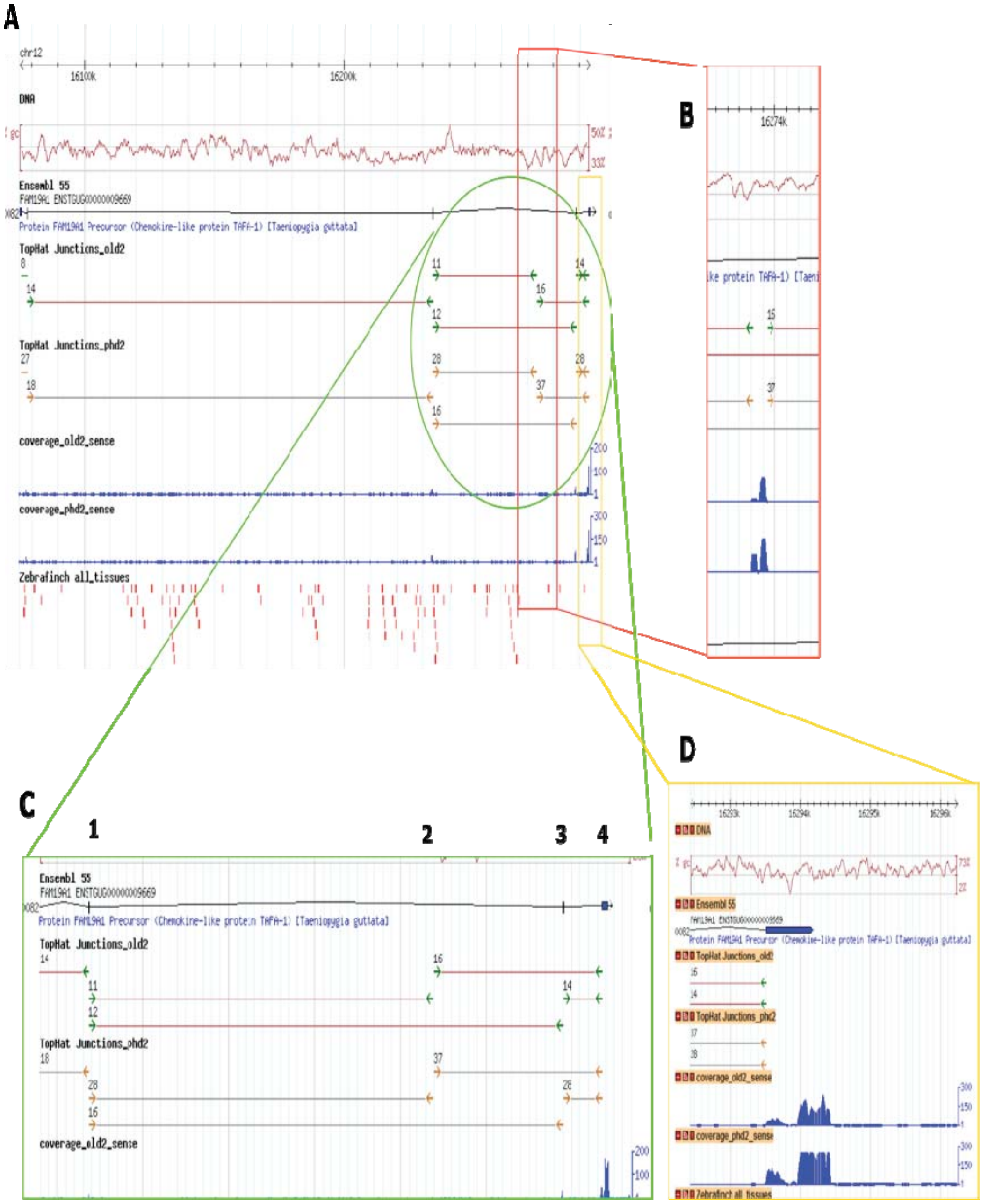
Supplementary Table 8. Statistical analysis of GO terms associated with song suppressible genes under positive selection (Supplementary Table 7).

GO	GO_description	total	exp	obs	fisher	adj. fisher
0005216	ion channel activity	49	1	6	1.50E-05	0.0016
0006811	ion transport	50	1	6	1.70E-05	0.0016
0005886	plasma membrane	95	1	7	8.00E-05	0.0052
0005249	voltage-gated potassium channel activity	34	0	4	0.00053	0.026

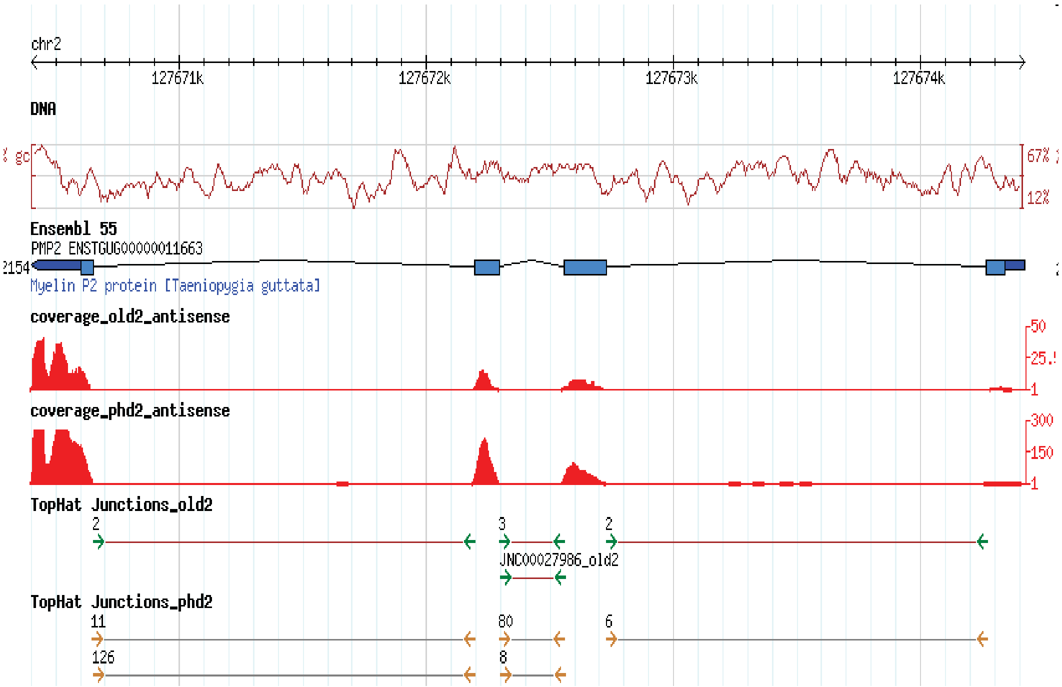
Supplementary Table 9. Genes annotated with the GO terms in Supplementary Table 8.

Gene	Associated Gene Name	Ion channel activity	Ion transport	Plasma membrane	Voltage-gated potassium channel activity
ENSTGUG00000002855	CACNA1B	•	•		•
ENSTGUG00000009049	CACNA1G	•	•	•	•
ENSTGUG00000000694	GPR98			•	
ENSTGUG00000005484	GRIA2	•	•	•	
ENSTGUG00000003563	GRIA3	•	•	•	
ENSTGUG00000007354	KCNC2	•	•		•
ENSTGUG00000008126	MOSPD2			•	
ENSTGUG00000005201	MYH10			•	
ENSTGUG00000007440	TRPM7-2			•	
ENSTGUG00000006952	TRPV1	•	•		
ENSTGUG00000011605	ZFHX4				•

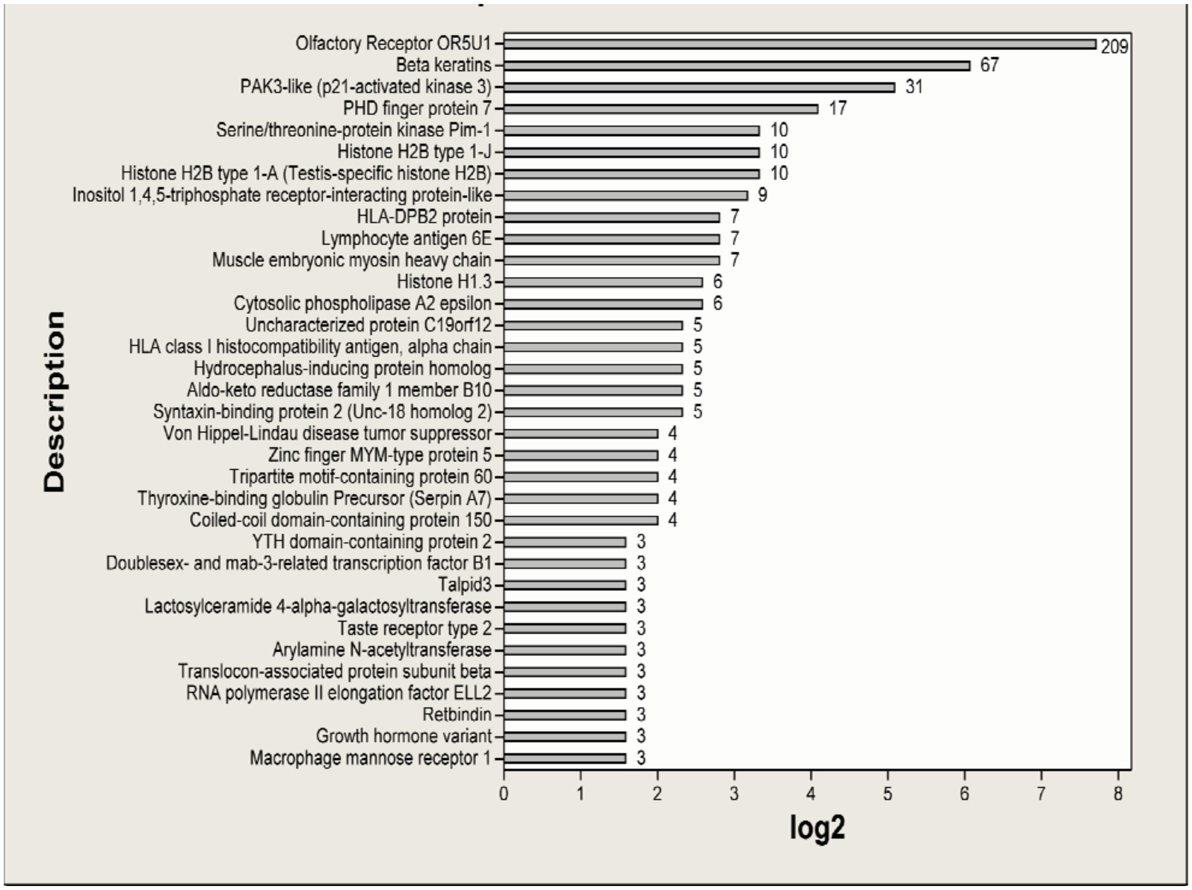
III. Supplementary figures



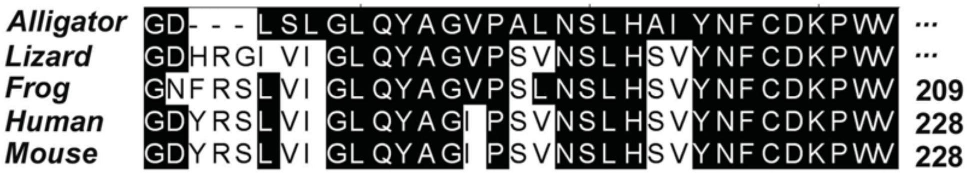
Supplementary Figure 1. Gene expression in the aging brain. A) Visualization of the gene FAM19A1 in a GMOD-based genome browser. The depth of sequencing as well as the random distribution of reads along the transcripts allows detection of new exons due to significant accumulation of reads in intronic regions (B). Evidence for the novel exon is furthermore strengthened by interrogation of reads spanning over introns, that directly support the integration of the novel exon into the transcript. These reads also allow an unprecedented view on the diversity of splice forms. (C) The example depicts mutually exclusive splicing, as according to the intron-spanning reads either exon 3 is retained in the transcript or exon 2 but no transcript variant seems to exist that includes both exons. (D) Finally the high coverage on the 3‘-end of the transcript enables refined annotation of 3‘UTRs and poly-adenylation signals.



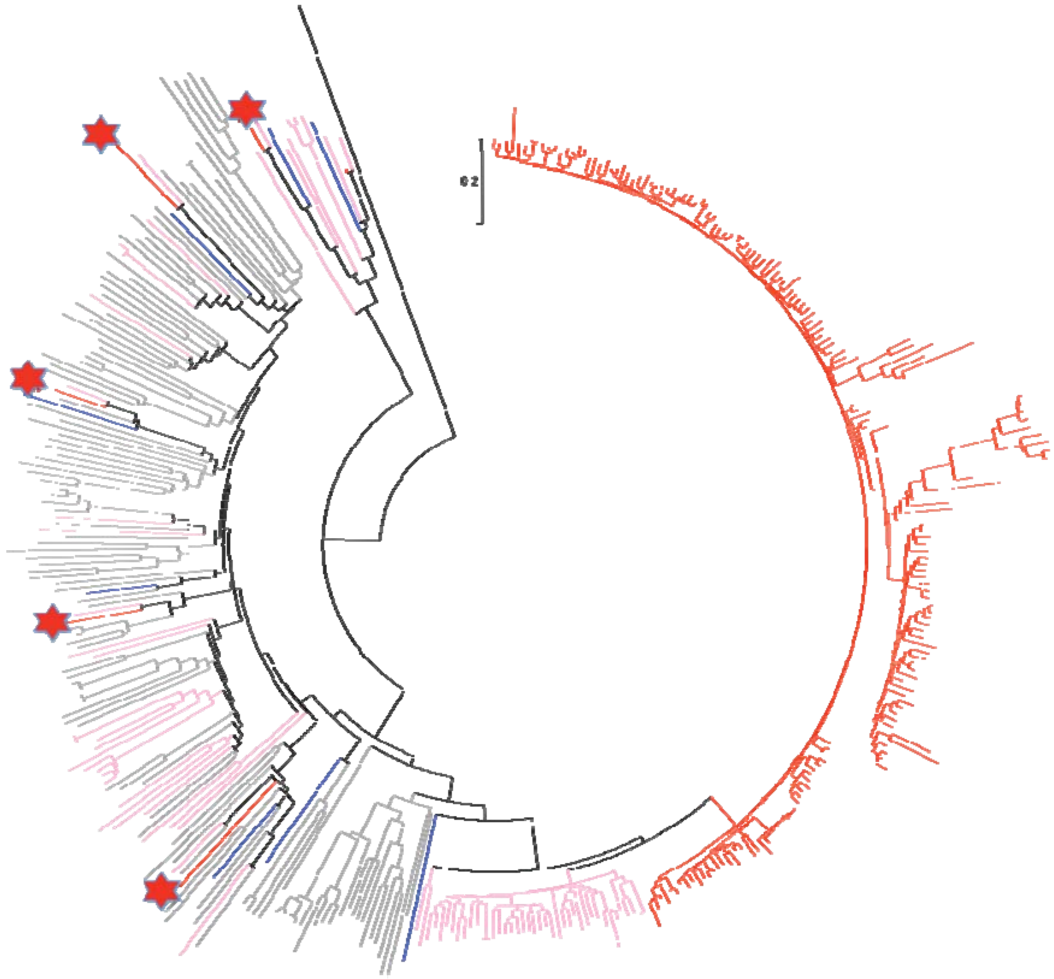
Supplementary Figure 2. **GBrowse visualization of PMP2**. One example of the 1,342 differentially expressed genes between the two ages analyzed. The gene product of PMP2 is implicated in myelination of nerve-cells, which is an ongoing process in young postnatal brain-tissue and much less so in adult tissue. This is reflected by the steep decrease of detected expression between the two ages (note the differences in scales).



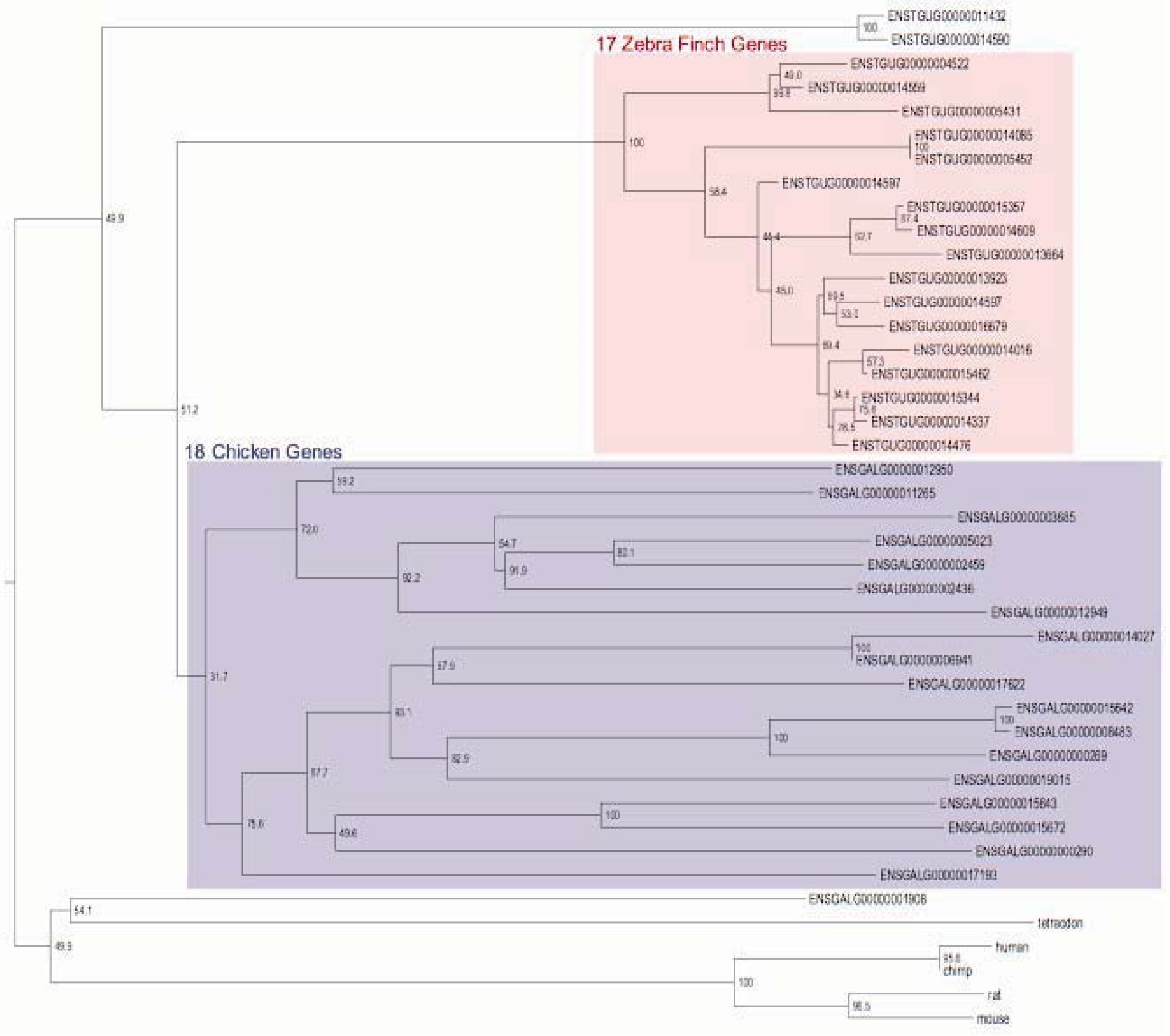
Supplementary Figure 3. **Gene expansions in the zebra finch sauropsid lineage, after the split with mammals, on a log2 scale**. For each group, the gene number in zebra finch was compared to the inferred gene number in the last common ancestor of amniotes. Genes within several groups (*OR5U1*, beta keratin, *PAK3*-like, *PHF7*, and *PIM1*-like) have been manually annotated. To avoid artifact duplications due to potential assembly errors, all apparently duplicated zebra finch genes on unplaced chromosome (chrUn) that were more that 97% sequence identical to their closest paralog were discarded, unless expression data were available for each duplicated gene.


Alligator GD- - - LSLGLQYAGVPALNSLHAI YNFCDKPW ...
Lizard GDHRGI VI GLQYAGVPSVNSLHSVYNFCDKPW ...
Frog GNFRSLVI GLQYAGVPSLNSLHSVYNFCDKPW 209
Human GDYRSLVI GLQYAGI PSVNSLHSVYNFCDKPW 228
Mouse GDYRSLVI GLQYAGI PSVNSLHSVYNFCDKPW 228

Supplementary Figure 4. ***SYN1* multispecies protein alignment**. The predicted amino acid sequence of a segment of the *Syn1* gene from the American alligator (*Alligator mississippiensis*) is aligned with lizard, frog, mouse and human orthologues (S2 supplementary notes). The alligator sequence was obtained through PCR amplification from genomic DNA (kindly provided by Dr. Travis Glenn at Savannah River Ecology Laboratory, University of Georgia) using primers designed based on highly conserved regions of the *Syn1* gene. The other sequences were derived from either genomic (lizard) or cDNA sequences (others species). At the amino acid level, the alligator *Syn1* sequence shows 80% identity with *Syn1* from frog and ~76% identity with lizard, human and mouse *Syn1*. In contrast, this fragment shows only 71% and 68% identity when compared with the closest *Syn2* and *Syn3* sequences, from frog and mouse respectively. A comparable band could not be amplified from genomic DNA of the zebra finch, consistent with the prediction from the synteny analysis that *Syn1* is lacking in birds.



Supplementary Figure 5. **The zebra finch OR repertoire in comparison to other sauropsid ORs.** Neighbor-joining phylogenetic tree containing functional OR genes from Zebra finch (red), chicken (pink) and lizard (grey). Representative mammalian ORs are also shown (blue). Rhodopsin is used as outgroup. Distances are calculated using the Poisson correction[78]. Zebra finch ORs which share 'simple' (one-to-one) orthologs with other vertebrates are highlighted with '*'.



Supplementary Figure 6. Gene tree analysis of *PHF7* gene expansions. We aligned amino acid sequences from chicken, zebra finch and five outgroup species using Multiple Alignment using Fast Fourier Transform (MAFFT). Protein Neighbor Joining (distance) and Bootstrap analysis were done through Phylip-3.69 to construct phylogenetic trees. Analyses were run with 1000 replicates, all of which were used to generate a consensus tree. Node labels show bootstrap values.

IV. Supplementary References

[1] Huang X, Y. S., Chinwalla AT, Hillier LW, Minx P, Mardis ER, Wilson RK. Application of a superword array in genome assembly. *Nucleic Acids Res*. **34**, 201-205, (2006).

[2] Hillier, L. W. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716, (2004).

[3] Wallis, J. W. *et al.* A physical map of the chicken genome. *Nature* **432**, 761-764, (2004).

[4] Replogle, K. *et al.* The Songbird Neurogenomics (SoNG) Initiative: community-based tools and strategies for study of brain gene function and evolution. *BMC Genomics* **9**, 131, (2008).

[5] Li, X. *et al.* Genomic resources for songbird research and their use in characterizing gene expression during brain development. *Proc Natl Acad Sci* USA **104**, 6834-6839, (2007).

[6] Wada, K. *et al.* A molecular neuroethological approach for identifying and characterizing a cascade of behaviorally regulated genes. *Proc Natl Acad Sci* USA **103**, 15212-15217, (2006).

[7] Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. *Genome Res* **8**, 175-185, (1998).

[8] Stapley, J., Birkhead, T. R., Burke, T. & Slate, J. A linkage map of the zebra finch Taeniopygia guttata provides new insights into avian genome evolution. *Genetics* **179**, 651-667, (2008).

[9] Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res* **11**, 1725-1729, (2001).

[10] Itoh, Y. & Arnold, A. P. Chromosomal polymorphism and comparative painting analysis in the zebra finch. *Chromosome Res* **13**, 47-56, (2005).

[11] Hubbard, T. J. *et al.* Ensembl 2009. *Nucleic Acids Res* 37, D690-697, (2009).

[12] Heger, A. & Ponting, C. P. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Res* **17**, 1837-1849, (2007).

[13] Heger, A. & Ponting, C. P. OPTIC: orthologous and paralogous transcripts in clades. *Nucleic Acids Res* **36**, D267-270, (2008).

[14] Warren, W. C. *et al.* Genome analysis of the platypus reveals unique sig*Nature*s of evolution. *Nature* **453**, 175-183, (2008).

[15] Burt, D. W. *et al.* The Chicken Gene Nomenclature Committee report. *BMC Genomics* **10** Suppl 2, S5, (2009).

[16] Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res*, (2009).

[17] Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, (2009).

[18] Benjamini, Y. H., Y. . Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289-300, (1995).

[19] Alaux, C. *et al.* Honey bee aggression supports a link between gene regulation and behavioral evolution. *Proc Natl Acad Sci* USA, (2009).

[20] Blanchette, M. *et al.* Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* **16**, 656-668, (2006).

[21] Stankiewicz, P. & Lupski, J. R. Molecular-evolutionary mechanisms for genomic disorders. *Curr Opin Genet Dev* **12**, 312-319, (2002).

[22] Griffin, D. K., Robertson, L. B., Tempest, H. G. & Skinner, B. M. The evolution of the avian genome as revealed by comparative molecular cytogenetics. Cytogenet *Genome Res* 117, 64-77, (2007).

[23] Burt, D. W. *et al.* The dynamics of chromosome evolution in birds and mammals. *Nature* 402, 411-413, (1999).

[24] Burt, D. W. Origin and evolution of avian microchromosomes. Cytogenet *Genome Res* 96, 97-112, (2002).

[25] Choudhuri, J. V., Schleiermacher, C., Kurtz, S. & Giegerich, R. GenAlyzer: interactive visualization of sequence similarities between entire genomes. *Bioinformatics* 20, 1964-1965, (2004).

[26] Crooijmans, R. P., Vrebalov, J., Dijkhof, R. J., van der Poel, J. J. & Groenen, M. A. Two-dimensional screening of the Wageningen chicken BAC library. *Mamm Genome* 11, 360-363, (2000).

[27] Griffin, D. K. *et al.* Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution. *BMC Genomics* 9, 168, (2008).

[28] Skinner, B. M. *et al.* Comparative genomics in chicken and Pekin duck using FISH mapping and microarray analysis. *BMC Genomics* 10, 357, (2009).

[29] Lichter, P. *et al.* High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science* **247**, 64-69, (1990).

[30] Abramoff, M. D., Magelhaes, P.J., Ram, S.J. Image Processing with ImageJ. *Biophotonics International* **11**, 36-42, (2004).

[31] Tomaszycki, M. L. *et al.* Sexual differentiation of the zebra finch song system: potential roles for sex chromosome genes. *BMC Neurosci* 10, 24, (2009).

[32] Itoh, Y. *et al.* Dosage compensation is less effective in birds than in mammals. *J Biol* **6**, 2, (2007).

[33] Melamed, E. & Arnold, A. P. The role of LINEs and CpG islands in dosage compensation on the chicken Z chromosome. *Chromosome Res*, (2009).

[34] Goodstadt, L. & Ponting, C. P. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol* **2**, e133, (2006).

[35] Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-3402, (1997).

[36] Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC *Bioinformatics* **5**, 113, (2004).

[37] Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**, 327-335, (2009).

[38] Hedges, S. B. The origin and evolution of model organisms. *Nat Rev Genet* **3**, 838-849, (2002).

[39] Sonnhammer, E. L. & Koonin, E. V. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* **18**, 619-620, (2002).

[40] Chen, Q. *et al.* Family-based association study of synapsin II and schizophrenia. *Am J Hum Genet* **75**, 873-877, (2004).

[41] Lachman, H. M., Stopkova, P., Rafael, M. A. & Saito, T. Association of schizophrenia in African Americans to polymorphism in synapsin III gene. *Psychiatr Genet* **15**, 127-132, (2005).

[42] Cavalleri, G. L. *et al.* Multicentre search for genetic susceptibility loci in sporadic epilepsy syndrome and seizure types: a case-control study. *Lancet Neurol* **6**, 970-980, (2007).

[43] Gitler, D. *et al.* Different presynaptic roles of synapsins at excitatory and inhibitory synapses. *J Neurosci* **24**, 11368-11380, (2004).

[44] Gaffield, M. A. & Betz, W. J. Synaptic vesicle mobility in mouse motor nerve terminals with and without synapsin. *J Neurosci* **27**, 13691-13700, (2007).

[45] Balthazart, J. & Taziaux, M. The underestimated role of olfaction in avian reproduction? *Behav Brain Res* **200**, 248-259, (2009).

[46] DeBose, J. L. & Nevitt, G. A. The use of odors at different spatial scales: comparing birds with fish. *J Chem Ecol* **34**, 867-881, (2008).

[47] Lagerstrom, M. C. *et al.* The G protein-coupled receptor subset of the chicken genome. *PLoS Comput Biol* **2**, e54, (2006).

[48] Niimura, Y. & Nei, M. Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *Proc Natl Acad Sci* USA **102**, 6039-6044, (2005).

[49] Steiger, S. S., Fidler, A. E., Valcu, M. & Kempenaers, B. Avian olfactory receptor gene repertoires: evidence for a well-developed sense of smell in birds? *Proc Biol Sci* **275**, 2309-2317, (2008).

[50] Quesada, V., Ordonez, G. R., Sanchez, L. M., Puente, X. S. & Lopez-Otin, C. The Degradome database: mammalian proteases and diseases of proteolysis. *Nucleic Acids Res* **37**, D239-243, (2009).

[51] Lopez-Otin, C. & Bond, J. S. Proteases: multifunctional enzymes in life and disease. *J Biol Chem* **283**, 30433-30437, (2008).

[52] Puente, X. S., Sanchez, L. M., Overall, C. M. & Lopez-Otin, C. Human and mouse proteases: a comparative genomic approach. *Nat Rev Genet* **4**, 544-558, (2003).

[53] Gertz, E. M., Yu, Y. K., Agarwala, R., Schaffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol* **4**, 41, (2006).

[54] Huesmann, G. R. & Clayton, D. F. Dynamic role of postsynaptic caspase-3 and BIRC4 in zebra finch song-response habituation. *Neuron* **52**, 1061-1072, (2006).

[55] Eckhart, L. *et al.* Identification of novel mammalian caspases reveals an important role of gene loss in shaping the human caspase repertoire. *Mol Biol Evol* **25**, 831-841, (2008).

[56] Molinari, F. *et al.* Truncating neurotrypsin mutation in autosomal recessive nonsyndromic mental retardation. *Science* **298**, 1779-1781, (2002).

[57] Didelot, G. *et al.* Tequila, a neurotrypsin ortholog, regulates long-term memory formation in Drosophila. *Science* **313**, 851-853, (2006).

[58] Tesco, G. *et al.* Depletion of GGA3 stabilizes BACE and enhances beta-secretase activity. *Neuron* **54**, 721-737, (2007).

[59] Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755, (2001).

[60] Balakrishnan, C. N., Ekblom, R., Völker, M., Westerdahl, H., Kotkiewicz, H., Godinez, R., Burt, D.W., Graves, T., Griffin, D.K., Warren, W.C., & Edwards, S.V. Gene duplication and fragmentation in the zebra finch Major Histocompatibility Complex *BMC Genomics*, under review (2010).

[61] Ekblom, R., Balakrishnan, C.N., Burke, T., & Slate, J. Digital gene expression analysis of the zebra finch genome and major histocompatibility complex. *BMC Genomics*, in press (2010).

[62] Gu, W., Castoe, T. A., Hedges, D. J., Batzer, M. A. & Pollock, D. D. Identification of repeat structure in large genomes using repeat probability clouds. *Anal Biochem* **380**, 77-83, (2008).

[63] Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* 11, 1005-1017, (2001).

[64] She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927-930, (2004).

[65] Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003-1007, (2002).

[66] Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859-1875, (2005).

[67] Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, (2000).

[68] Dong, S. *et al.* Discrete molecular states in the brain accompany changing responses to a vocal signal. *Proc Natl Acad Sci* USA **106**, 11364-11369, (2009).

[69] Kong, L. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**, W345-349, (2007).

[70] Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282-1288, (2007).

[71] Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**, 556-565, (2007).

[72] Spiteri, E. *et al.* Identification of the transcriptional targets of FOXP2, a gene linked to speech and language, in developing human brain. *Am J Hum Genet* **81**, 1144-1157, (2007).

[73] Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80, (2004).

[74] Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18** Suppl 1, S96-104, (2002).

[75] Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3, (2004).

[76] Jarvis, E. D. & Nottebohm, F. Motor-driven gene expression. *Proc Natl Acad Sci* USA **94**, 4097-4102, (1997).

[77] Vlieghe, D. *et al.* A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* **34**, D95-97, (2006).

[78] Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci* USA **100**, 9440-9445, (2003).

[79] Wingender, E. *et al.* TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**, 316-319, (2000).

[80] Flavell, S. W. & Greenberg, M. E. Signaling mechanisms linking neuronal activity to gene expression and plasticity of the nervous system. *Annu Rev Neurosci* **31**, 563-590, (2008).

[81] Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591, (2007).

[82] Wu, X. & Watson, M. CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics* **25**, 832-833, (2009).

[83] Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-556, (1997).

[84] Yang, Z., Wong, W. S. & Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**, 1107-1118, (2005).

[85] Nam, K. *et al.* Comparative genomics and gene sequence evolution in birds. (submitted).